Dpto. de Teoría de la Señal y Comunicaciones
Universidad Carlos iii de Madrid

TESIS DOCTORAL

CONTRIBUCIONES AL RECONOCIMIENTO ROBUSTO DE HABLA

Autor: Jesús de Vicente Peña
Director: Dr. Fernando Díaz De María

Leganés, 2007

Tesis Doctoral:
CONTRIBUCIONES AL RECONOCIMIENTO ROBUSTO DE HABLA

Autor:
JESÚS de VICENTE PEÑA

Director:
Dr. FERNANDO DÍAZ DE MARÍA

Firma del Tribunal Calificador:

Presidente:

Vocal:

Vocal:

Vocal:

Secretario:

Calificación:

Leganés, de de

RESUMEN

En esta tesis doctoral abordamos el problema del reconocimiento robusto de habla. En primer lugar, estudiamos el efecto de los ruidos aditivos sobre el proceso de reconocimiento. Mostramos que el deterioro de la eficacia de los reconocedores se debe, en parte, al excesivo poder de decisión que recae sobre características de entrada que están contaminadas de forma muy severa (*outliers*). El método que denominamos *bounded-distance* HMM (BD-HMM) es capaz de limitar la contribución de estas muestras en el reconocedor. Sin embargo, BD-HMM no actúa sobre el resto de observaciones que, sin estar tan altamente contaminadas, sí que están afectadas por la presencia de ruidos aditivos. Por el contrario, sustracción espectral actúa sobre todas las observaciones pero presenta el inconveniente de introducir distorsiones que afectan negativamente a las prestaciones de los reconocedores. En esta tesis mostramos que estas distorsiones producen un aumento del número de *outliers*. De este modo, encontramos que BD-HMM y sustracción espectral se complementan perfectamente. Nuestros experimentos muestran que esta combinación generalmente supera las tasas de reconocimiento que consiguen tanto BD-HMM como sustracción espectral cuando se aplican de forma aislada. De hecho, las mejoras introducidas por la combinación propuesta, especialmente a SNRs bajas y medias, suelen ser superiores a la suma de las mejoras conseguidas por BD-HMM y sustracción espectral.

Por otro lado, la estimación de los parámetros limpios que realiza sustracción espectral deja un cierto nivel de incertidumbre que los reconocedores convencionales no tienen en consideración. En esta tesis modificamos el proceso de reconocimiento para tener en cuenta esta incertidumbre cuando los sistemas se diseñan empleando la parametrización *Frequency Filtered* (FF). Al permanecer esta parametrización en el dominio del log-espectro, los métodos propuestos se pueden interpretar de una forma sencilla como métodos de ponderación espectral que asignan mayor poder discriminativo a las frecuencias del espectro más fiables. Los resultados que mostramos en esta tesis apoyan la necesidad de incorporar información sobre la incertidumbre de las observaciones para aumentar la robustez del proceso de reconocimiento.

Por último, en esta tesis abordamos el problema del reconocimiento de habla cuando la señal de voz es transmitida a través de un canal inalámbrico. Las distorsiones que este tipo de entornos introducen en los parámetros de entrada son más difíciles de modelar que en el caso de tener ruidos aditivos y, por ello, su efecto se ha estudiado de forma experimental en el dominio del espectro de modulación. A la vista de nuestras observaciones, proponemos filtrar paso-banda la evolución temporal de los parámetros para aumentar la robustez del sistema reconocedor. Nuestra propuesta se evalúa para dos parametrizaciones bajo canales con diferentes tasas de error de bit (*Bit Error Rate*, BER) típicas de este tipo de comunicaciones inalámbricas: por un lado, filtramos paso-banda la evolución temporal de los parámetros LP-MCC y, por otro, sustituimos el filtro paso-banda RASTA-PLP por otro cuya sección paso bajo es más abrupta. Nuestros resultados encuentran mejores resultados con las secuencias filtradas. Finalmente, aplicamos la técnica BD-HMM para reducir el impacto de los *outliers* en este tipo de entornos inalámbricos. Nuestros resultados muestran que BD-HMM introduce importantes mejoras para canales con altas tasas de error de bit.

ABSTRACT

In this Ph.D. Thesis we address the problem of robust speech recognition. We start studying the effects of additive noises. We show that one of the causes contributing to the loss of performance in presence of noise is the fact that conventional recogniser take into consideration feature values that are actually outliers. We propose a method that we call Bounded-Distance HMM (BD-HMM) to mitigate the outlier contribution to the recogniser decision.

Since BD-HMM just deals with outliers, leaving the remaining features unaltered, we suggest to combine it with other techniques that work on all the features. In particular, we propose to use spectral subtraction as feature enhancement technique, since it complements BD-HMM well. As we prove in the Thesis, spectral subtraction introduces some artifacts that cause a larger number of outliers that can be easily countered by BD-HMM. Our experimental results show that the combination of these techniques generally outperforms both BD-HMM and spectral subtraction individually. Furthermore, the obtained improvements, especially for low and medium SNRs, are generally larger than the sum of the improvements individually obtained by BD-HMM and spectral subtraction.

On the other hand, the spectral subtraction-based estimates of the original parameters generate certain level of uncertainty that is not usually taken into account by the decoding algorithm. This Thesis takes into consideration this uncertainty in the recogniser for a specific type of features: the Frequency Filtered parameterization. Moreover, as this parameterization remains in the log-frequency domain, the proposed method admits a simple interpretation as a spectral weighting method that assigns more importance to the most reliable spectral components. Our results show the convenience of incorporating this information in the decoding process.

Finally, in this Thesis we tackle the problem of speech recognition when wireless speech communication systems are involved. The distortions caused by this environment are more difficult to model analytically than the ones caused by additive noises. Thus, we experimentally study their effects on the feature spectra and we

propose to band-pass filter the recognition features to improve the ASR performance. We have evaluated our proposal in two configurations at different Bit Error Rates (BER) typical of these channels: band-pass filtering the LP-MFCC parameters and a modification of the RASTA-PLP using a sharper low-pass section. Both filtered parameterizations perform consistently better than the unfiltered ones. Additionally, we remove the impact of the outliers by applying BD-HMM, what results in larger improvements for high BER channels.

AGRADECIMIENTOS

En primer lugar, deseo expresar mi más sincero agradecimiento al Dr. Fernando Díaz de María, director de esta tesis. Debo destacar su profesionalidad y agradecer la paciencia, comprensión y apoyo que me ha dedicado durante todos estos años. Siempre admiraré su gran capacidad de dirección y optimismo.

En segundo lugar, estoy especialmente agradecido al Prof. Dr. Bastiaan Kleijn, que me acogió en su laboratorio en KTH (Estocolmo) durante 10 meses y al que considero mi segundo director de tesis. Las propuestas realizadas en los Capítulos 3 y 4 no hubieran sido posibles sin su estrecha colaboración. Ha sido un gran honor trabajar con él y aprender de su profesionalidad y experiencia.

En tercer lugar, agradezco al Prof. Dr. Richard Stern que me permitiese visitar su laboratorio en CMU (Pittsburgh) durante 2 meses en los que tanto disfruté empapándome de su experiencia en el campo del reconocimiento robusto de habla.

También quiero reconocer a las Dras. Ascensión Gallardo Antolín y Carmen Peláez Moreno su colaboración en las propuestas realizadas en el Capítulo 5. Esta tesis no hubiera sido posible sin el trabajo realizado por las personas encargadas del mantenimiento de las infraestructuras informáticas, en especial, agradezco al Dr. Harold Molina Bulla su enorme dedicación al mantenimiento de la granja de computación y al Ing. Saúl Blanco Fortes su trabajo en el soporte informático.

Quiero realizar una mención especial a mis compañeros de trabajo empezando por los que me encontré en aquel verano en CMU, siguiendo por los que tan acogedoramente me recibieron en la universidad de KTH y terminando por los que han compartido más tiempo conmigo en el departamento. Además, debo destacar a todos aquellos que han hecho que este periodo de tiempo haya sido tan agradable: a mis compañeros de comedor, a los que fueron mis compañeros de laboratorio y a mi compañero de despacho, Emilio, que aguanta y atiende tan pacientemente las dudas que a diario le planteo. En este sentido, agradezco de nuevo a Asen todos los consejos que me ha dado durante todos estos años.

Tampoco puedo olvidar a mis amigos de fuera de la universidad: los que aguantan

desde el instituto; los incondicionales de la carrera; los que conocí a través de mi gran vecino Jorge; los de Pittsburgh y de Estocolmo; los de mi equipo de fútbol; a la familia Andersson ...

A mi familia, por su cariño, dedicación y apoyo incondicional y que representan, de una manera u otra, todos los valores y virtudes que considero importantes. A mis hermanos: Yolanda, Belén, JuanMa y Carlos; a mis cuñados Guillermo y Mario; mis sobrinos: Celia, Inés, Pablo y Diego; a Maria, por su apoyo, cariño y querer formar parte de mi vida. Y, por último, a mis padres, Juan Manuel e Inés, claros referentes en mi vida, esta tesis va especialmente dedicada a ellos.

Jesús de Vicente Peña
Madrid, diciembre 2007.

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor Dr. Fernando Díaz-de-María. In addition to his professionalism, I greatly appreciate his patience, understanding and support. I will always admire his direction skills and his optimism.

Second, I am especially grateful to Prof. Dr. Bastiaan Kleijn who took me in his laboratory at KTH (Stockholm) during 10 months and who I think of as my second advisor. The proposals of Chapters 3 and 4 would have not been possible without his close collaboration. It has been a great honour working with him and learning from his professionalism and experience.

Third, I thank Prof. Dr. Richard Stern for giving me the opportunity of visiting his laboratory at CMU (Pittsburgh) for 2 months, where I enjoyed so much learning from his experience in the field of robust speech recognition.

In addition, I thank Dr. Ascensión Gallardo-Antolín and Dr. Carmen Peláez-Moreno for their collaboration in the contents of Chapter 5. This Ph.D. Thesis would have not been possible without the people in charge of the software and hardware infrastructures: I thank Dr. Harold Molina-Bulla for his enormous dedication to the compute farm maintenance and Ing. Saúl Blanco-Fortes for his IT support.

Thanks to my great workmates: to the ones that I met at CMU on that summer; to those who welcomed me so kindly at KTH and, finally, to the guys that have spent more time with me at the department. Furthermore, thanks to the people that made my time much nicer and enjoyable, specially, to my lunch mates, to my former lab mates and to my office mate, Emilio, for answering with such patience my endless technical and non-technical questions. In this sense, I wish to thank again Asen for her invaluable advises during my Ph.D.

Thanks to all my friends from outside university: thanks to who keep on since high-school; to my unconditional friends from my undergraduate degree; to the ones who I know from my great neighbour Jorge; to my friends in Pittsburgh and Stockholm; to my football team; to the Anderssons; ...

My thanks to my family, for their love and unconditional support and who repre-

sent all the values and virtues worth admiring. To my sisters and brothers: Yolanda, Belén, JuanMa and Carlos; to my brothers-in-law Guillermo and Mario; to my nieces and nephews: Celia, Inés, Pablo and Diego; to Maria, for her support, love and for being part of my life. And, finally, to my parents, Juan Manuel e Ines, undoubtedly my truly role models, this Ph.D. Thesis is dedicated to them.

Jesús Vicente-Peña

Madrid, December 2007.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En este primer capítulo se describe el problema de reconocimiento de habla y se exponen la motivación y los objetivos de esta tesis, para terminar describiendo la estructura que se sigue en el resto del documento.

1.1. Reconocimiento automático de habla

El objetivo de un sistema de reconocimiento automático de habla (RAH) es extraer la información lingüística contenida en la señal de voz. Para ello determina la secuencia de palabras más verosímil dada la señal de voz de entrada. Dicho criterio puede expresarse matemáticamente de la siguiente manera:

$$\mathbf{W} = \arg\max_{i}\{P(\mathbf{W_i}|\mathbf{O})\} \tag{1.1}$$

siendo $\mathbf{W_i}$ la hipótesis i sobre la secuencia de palabras y $\mathbf{O}$ la secuencia de observaciones obtenida a partir de la señal de voz. Esta secuencia de observaciones no será más que la secuencia de vectores de parámetros que representan la señal de voz y que se obtienen para cada trama:

$$\mathbf{O} = \{\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_T}\} \tag{1.2}$$

donde $\mathbf{o_t}$ denota el vector de parametrización obtenido para el instante t, y T indica el número de vectores de parámetros que extraemos de la señal observada.

Aplicando la regla de Bayes a la ecuación (1.1) obtenemos la siguiente expresión alternativa:

$$\mathbf{W} = \arg\max_i \left\{ \frac{P(\mathbf{O}|\mathbf{W_i})P(\mathbf{W_i})}{P(\mathbf{O})} \right\}, \tag{1.3}$$

y habida cuenta de que la probabilidad de las observaciones, $P(\mathbf{O})$, es independiente de la hipótesis considerada, es posible prescindir de este término en la maximización:

$$\mathbf{W} = \arg\max_i \{P(\mathbf{O}|\mathbf{W_i})P(\mathbf{W_i})\} \tag{1.4}$$

En esta nueva expresión se diferencian claramente dos términos: el primero, $P(\mathbf{O}|\mathbf{W_i})$, indica cuánto de verosímil es la señal de voz observada dada la hipótesis i; el segundo, $P(\mathbf{W_i})$, representa la probabilidad a priori de dicha hipótesis o, dicho de otro modo, la probabilidad de observar la secuencia de palabras representada por la hipótesis i. Estos términos se obtienen evaluando, respectivamente, lo que se conoce como el modelo acústico y el modelo de lenguaje propios del sistema de reconocimiento; ambos modelos constituyen los dos pilares básicos sobre los que se construye el proceso de reconocimiento automático de habla.

En la Figura 1.1 se representan esquemáticamente los elementos principales de un sistema de reconocimiento. En las subsecciones siguientes se explica cada bloque del diagrama de la Figura 1.1 con cierto detalle. Así, en la subsección siguiente se describen las parametrizaciones más habituales para representar la señal de voz; a continuación, se explica el modelado acústico, el de lenguaje y, finalmente, el proceso que nos permite obtener la transcripción del mensaje (la decodificación).

1.1.1. Parametrización

Modelar directamente la forma de onda de la señal de voz no resulta adecuado para discriminar unos sonidos frente a otros y, por tanto, resulta necesario obtener otros parámetros.

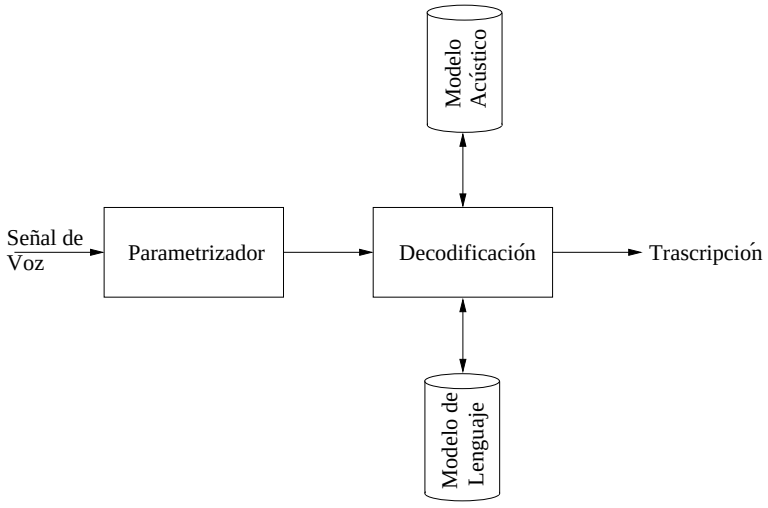


Figura 1.1: Elementos principales de un sistema de reconocimiento automático de habla.

El modelo de producción de voz que se emplea habitualmente contempla dos componentes en la señal de voz. Por un lado encontramos la excitación, que es responsable de la estructura fina del espectro y cuya utilidad es limitada a la hora de diferenciar unos sonidos frente a otros. Por otro lado, encontramos la componente que relacionamos con la articulación de los sonidos; la articulación determina la forma de la envolvente espectral de la señal de voz y, por tanto, una buena parametrización debe ser capaz de estimar dicha envolvente a la vez que no se ve afectada por los detalles que porta la excitación. Además, una buena parametrización debe ajustarse al sistema de reconocimiento de modo que a partir de ella puedan construirse modelos estadísticos eficaces de cada sonido (modelos acústicos). En el resto de esta subsección se expone con más detalle el procesado al que se somete generalmente la señal de voz para representar adecuadamente los diferentes sonidos.

Análisis de la señal de voz

En la Figura 1.2 hemos representado gráficamente la metodología que se sigue para realizar el análisis de la señal de voz. En dicha figura vemos como, por medio

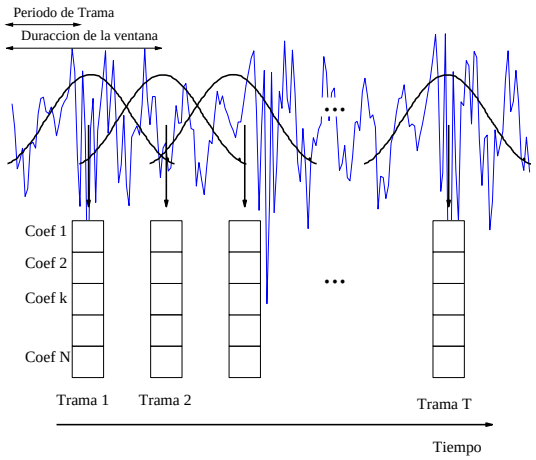


Figura 1.2: Análisis de la señal de voz: enventanado

de ventanas, se seleccionan segmentos temporales a partir de los que obtenemos los parámetros de interés. La elección de la ventana involucra un compromiso entre tres factores [O'Shaughnessy, 1999]:

- debe ser suficientemente corta como para que las propiedades de interés de la señal de voz no cambien;
- debe ser suficientemente larga como para que las medidas realizadas sobre esa ventana sean fiables; y por último,
- dos ventanas consecutivas deben estar lo suficientemente próximas para que las transiciones rápidas de la señal de voz se puedan medir correctamente.

Las dos primeras condiciones imponen restricciones relativas a la duración de la ventana mientras que la tercera se resuelve introduciendo cierto solapamiento entre ventanas adyacentes. Así lo vemos en la Figura 1.2, donde el parámetro "periodo de trama" representa el tiempo transcurrido entre dos medidas, que es menor que la duración de la ventana. Además, es común atenuar los extremos de las ventanas de modo que las medidas no fluctúen en exceso de una ventana a otra.

Siguiendo estas directrices, a lo largo de esta tesis se han empleado ventanas Hamming de 25 ms de duración cada 10 ms.

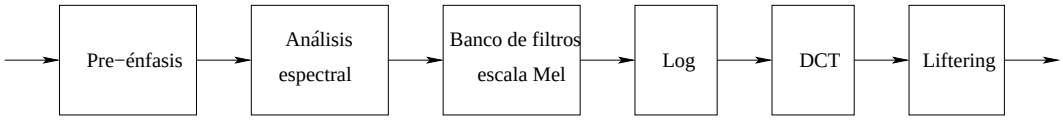


Figura 1.3: Diagrama de bloques ilustrativo del proceso de obtención de la parametrización MFCC

En el resto de esta subsección hacemos un breve repaso de los tipos de parámetros más habituales en reconocimiento de habla. Estos parámetros, como hemos indicado anteriormente, se calculan para cada ventana con el propósito de extraer características que permitan diferenciar unos sonidos frente a otros.

MFCC (Mel-Frequency Cepstral Coefficients)

La parametrización MFCC [Davis and Mermelstein, 1980] es posiblemente la parametrización más popular en los sistemas de reconocimiento automático de habla actuales. En la Figura 1.3 presentamos las etapas involucradas en la obtención de estos parámetros.

En el esquema representado en dicha figura partimos de la señal de voz enventanada y obtenemos, para cada ventana, un conjunto de parámetros MFCC. Cada uno de los pasos involucrados en la obtención de los parámetros MFCC tiene un objetivo que relacionamos o bien con características perceptuales de nuestro sistema auditivo o bien con la adecuación de los parámetros al sistema de reconocimiento. A continuación describimos brevemente el objetivo de cada una de estas etapas:

- Pre-énfasis: el espectro típico de los sonidos sonoros decae al aumentar la frecuencia y aplicamos este filtro para compensar esta pendiente en el espectro. Este paso también lo podemos considerar heredado del análisis mediante predicción lineal (conocido como análisis LP: *linear Prediction*) al que se suele someter la señal de voz; sin esta etapa de prepocesado no seríamos capaces de modelar correctamente los formantes situados en las frecuencias superiores.

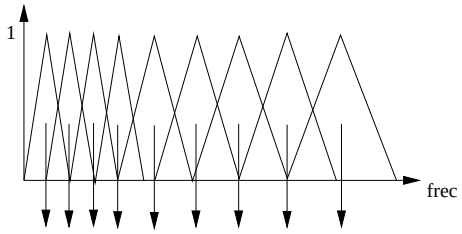


Figura 1.4: Representación simplificada del banco de filtros según la escala Mel.

Concretamente, la transformación que se realiza a la señal de voz es la siguiente:

$$s'_n = s_n - \kappa s_{n-1} \tag{1.5}$$

donde s_n y s'_n representan, respectivamente, la muestra de voz en el instante n antes y después del filtro de pre-énfasis; por último, κ es el coeficiente de pre-énfasis. Un valor habitual para este coeficiente es 0.97, es decir, estamos aplicando un filtro paso-alto a la señal de voz. Por último, pese a que en el esquema de la Figura 1.3 no se ha representado explícitamente el enventanado de la señal de voz, es habitual que este enventanado se realice tras la etapa de pre-énfasis.

- Análisis espectral: se calcula el espectro a corto plazo realizando la transformada de Fourier de la señal de voz enventanada. Además, ya que la fase no aporta información relevante para la discriminación de los sonidos, se calcula o bien el módulo del espectro o bien su densidad espectral de potencia.

- Banco de filtros escala Mel: en esta etapa se aplica un banco de filtros al módulo (o a la potencia) del espectro obtenido en la etapa anterior. A través de este banco de filtros simulamos el comportamiento del sistema auditivo, ya que estos filtros están distribuidos de forma no uniforme a lo largo del eje de frecuencias aportando mayor resolución a bajas frecuencias que a altas. En la Figura 1.4 hemos representado de forma gráfica y simplificada este banco de filtros cuya forma suele ser triangular.

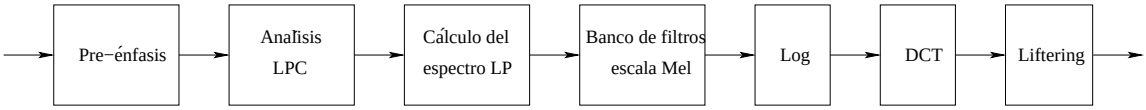


Figura 1.5: Diagrama de bloques ilustrativo del proceso de obtención de la parametrización LP-MFCC

- Log (Logaritmo): a través de este operador no lineal la señal de voz se divide en sus dos componentes principales: la excitación y la envolvente espectral. Además, este logaritmo permite simular el comportamiento del oído y su sensibilidad ante distintas intensidades de presión sonora.

- DCT (*Discrete Cosine Transform*): las log-energías en banda obtenidas en la etapa anterior están altamente correlacionadas y por medio de esta transformación conseguimos coeficientes más incorrelacionados. Además, la DCT concentra en los coeficientes bajos las componentes que provienen de la envolvente espectral y en los altos los que provienen de la excitación.

- Liftering: mediante esta etapa seleccionamos los coeficientes que representan a la envolvente espectral y eliminamos los representativos de la excitación.

LP-MFCC (Linear Prediction-MFCC)

La parametrización LP-MFCC se diferencia de la parametrización MFCC en la manera de estimar el espectro de la señal de voz. En la sección anterior vimos que los parámetros MFCC se obtenían a partir del espectro de la señal de voz calculado mediante una transformada de Fourier. En la parametrización LP-MFCC ese espectro se estima a partir del modelo todo-polos. Tras la estimación de este modelo, se calcula el módulo o la densidad espectral de potencia asociada al mismo. El esquema que describe los pasos para obtener la parametrización LP-MFCC queda ilustrado en la Figura 1.5.

Comparando las Figuras 1.3 y 1.5 vemos que, respecto a la parametrización MFCC, LP-MFCC difiere en la etapa de análisis espectral que ahora se compone

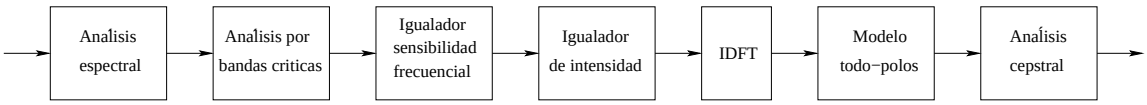


Figura 1.6: Diagrama de bloques ilustrativo del proceso de obtención de la parametrización PLP

de un análisis LPC y del cálculo del espectro LP. A partir del espectro LP todas las etapas coinciden con las correspondientes en la parametrización MFCC.

PLP (Perceptually-based linear prediction)

Otra parametrización ampliamente utilizada en los sistemas de reconocimiento de habla es la parametrización PLP. Aunque con sus particularidades, esta parametrización presenta bastantes aspectos comunes a las parametrizaciones MFCC y LP-MFCC.

En la Figura 1.6 presentamos el esquema de bloques que ilustra el cálculo de los coeficientes PLP. En lo que resta de sección, detallamos la funcionalidad de cada uno de estos bloques y los comparamos con las correspondientes etapas involucradas en el cálculo de las parametrizaciones MFCC y LP-MFCC. Un análisis similar puede encontrarse en [Gold and Morgan, 2000]:

- Análisis espectral: al igual que en el análisis MFCC, se calcula el módulo del espectro o su densidad espectral de potencia por medio de la transformada de Fourier.

- Análisis por bandas críticas: podemos considerar esta etapa equivalente a la fase donde se aplica el banco de filtros Mel en las otras parametrizaciones. La única diferencia destacable es que ahora este banco de filtros es trapezoidal en lugar de triangular. Aún así, el objetivo sigue siendo el mismo: simular la diferente sensibilidad del oído a altas y bajas frecuencias.

- Igualador sensibilidad frecuencial: esta fase es comparable a la fase de pre-énfasis en las otras parametrizaciones. Su objetivo es realzar las altas frecuencias

tal y como hace el sistema auditivo.

- Igualador de intensidad: esta etapa se incluye con el objetivo de simular la relación entre intensidad y tonalidad (intensidad percibida). Esta etapa se realiza a través del logaritmo en las otras parametrizaciones, mientras que aquí se emplea el operador raíz cúbica.

- Modelo todo-polos: tras volver al dominio del tiempo, se estima la envolvente espectral mediante un modelo todo-polos. Esta etapa se realiza en la parametrización LP-MFCC antes de incluir todos los efectos perceptuales que considera PLP.

- Análisis cepstral: por último, se obtienen componentes decorrelacionadas que se adaptan mejor a los sistemas de reconocimiento. Adicionalmente, se seleccionan los coeficientes de menor orden a través de un liftering.

Como hemos visto, la parametrización PLP presenta características comunes a la parametrización MFCC y LP-MFCC. Todas ellas reducen la resolución a altas frecuencias por medio de un banco de filtros y además proporcionan salidas decorrelacionadas que se adaptan mejor a los sistemas de reconocimiento.

La principal diferencia entre unas y otras radica en la naturaleza del suavizado espectral que se lleva a cabo. En las parametrizaciones PLP y LP-MFCC este suavizado se lleva a cabo estimando los polos de la envolvente espectral mientras que, en la parametrización MFCC, es el liftering el que realiza el suavizado.

Parámetro de log-energía

Como hemos visto anteriormente, cada trama de la señal de voz viene caracterizada por un vector de parámetros que representa la envolvente espectral de dicha trama. El objetivo de estos vectores a la entrada del reconocedor es distinguir unos sonidos frente a otros. Otra característica de la trama que es de gran utilidad para este objetivo es su energía. De este modo, las características estáticas a la entrada del reconocedor se completan utilizando este parámetro.

En general, en lugar de añadir el parámetro energía directamente se añade su logaritmo:

$$E = \log \sum_n s_n^2 \tag{1.6}$$

donde s_n representa la muestra de la señal de voz en el instante n. La energía se calcula sobre la ventana actual y en la implementación utilizada a lo largo de esta tesis se calcula antes de aplicar la ventana Hamming y el pre-énfasis. Además, este parámetro se suele normalizar de modo que, a lo largo de una frase, el máximo valor sea 1. También se limita el rango dinámico saturando los valores que caigan por debajo de un cierto valor mínimo (en nuestros experimentos el rango dinámico dentro de una frase se iguala a 50 dB).

Parámetros dinámicos

Varios estudios perceptuales han mostrado que la información fonética más relevante de cada sílaba está contenida en las variaciones temporales del espectro. Es por ello que se propone complementar los vectores de parametrización con coeficientes que den cuenta de tales variaciones temporales. Estas características que se añaden reciben el nombre de parámetros dinámicos o de regresión [Furui, 1986] y mejoran significativamente las prestaciones de los sistemas de reconocimiento. Además, los parámetros dinámicos son más robustos ante distorsiones en la señal de voz y ayudan a reducir los efectos de un cambio de entorno acústico entre las fases de entrenamiento y reconocimiento [Junqua and Haton, 1995].

Además de medir la evolución temporal de la envolvente espectral, también es conveniente calcular los parámetros dinámicos sobre la log-energía.

Los parámetros de regresión o dinámicos se calculan a partir de los parámetros estáticos del siguiente modo:

$$\Delta o_{t_k} = \frac{\sum_{\theta=1}^{\Theta} \theta [o_{[t+\theta]_k} - o_{[t-\theta]_k}]}{2 \sum_{\theta=1}^{\Theta} \theta^2} \tag{1.7}$$

donde Δo_{t_k} denota al parámetro de regresión de primer orden o parámetro delta asociado al k-ésimo coeficiente del vector de parametrización en el instante de tiempo

t, o_{t_k}; Θ es un parámetro de diseño que delimita la ventana de actuación de los parámetros delta. En los experimentos realizados en esta tesis hemos usado una ventana temporal que abarca dos muestras pasadas y dos futuras; así, Θ toma el valor 2, lo que hace que los parámetros delta se calculen sobre 65 ms de la señal de voz (usando un periodo de trama de 10 ms y un tamaño de ventana de 25 ms).

Además de los parámetros de regresión de primer orden, los sistemas de reconocimiento suelen añadir también los de segundo orden. Estos simplemente se calculan aplicando la ecuación (1.7) a los parámetros delta. Si bien parece claro que incorporar los parámetros dinámicos de primer orden mejora las prestaciones de los reconocedores, las ventajas de añadir los coeficientes de segundo orden no son tan claras [Junqua and Haton, 1995].

1.1.2. Modelo acústico

Los modelos acústicos almacenados en el reconocedor permiten evaluar el término $P(\mathbf{O}|\mathbf{W_i})$ de la ecuación (1.4). Actualmente, la tecnología más extendida para abordar el cálculo de $P(\mathbf{O}|\mathbf{W_i})$ es la basada en modelos ocultos de Markov (*Hidden Markov Models* (HMMs)). En el resto de secciones que componen este breve repaso sobre reconocimiento de habla explicamos los principios de los HMMs. Esta descripción no pretende ser exhaustiva y remitimos al lector interesado a cualquiera de las múltiples referencias existentes en la literatura, como por ejemplo [Rabiner, 1989], [Young et al., 2002] o [Jurafsky and Martin, 2000].

Un HMM es una máquina de estados finitos donde en cada instante de tiempo

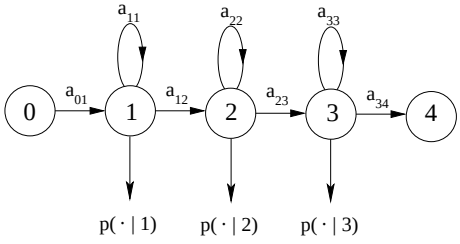


Figura 1.7: Topología ejemplo de un HMM

se produce un cambio de estado que está gobernado de manera probabilística. En la Figura 1.7 hemos representado un ejemplo de HMM. En este caso se trata de un HMM de 5 estados, siendo a_{lj} la probabilidad de pasar del estado l al j y $p(\cdot|j)$ la distribución de probabilidad que caracteriza la emisión de observaciones en el estado j. Los estados inicial y final son estados que se usan como meras conexiones cuando combinamos varios HMMs en cascada y, por ello, no emiten observaciones.

Apoyándonos en estos modelos paramétricos, $P(\mathbf{O}|\mathbf{W_i})$ se calcula como la probabilidad de que el HMM asociado a $\mathbf{W_i}$ haya generado la secuencia de observaciones $\mathbf{O}$. Ese HMM, que denotamos λ_i, se construye simplemente concatenando los HMMs que describen cada una de las palabras que componen la secuencia $\mathbf{W_i}$. A su vez, el HMM de cada palabra se compone de otros que describen unidades acústicas de nivel inferior. Estas unidades acústicas podrían ser fonemas o, lo que es más común, trifonemas, que ya tienen en cuenta información de contexto (efecto de coarticulación).

Atendiendo a la Figura 1.7, podemos calcular el término $p(\mathbf{O}|\lambda_i)$, que sirve como estimación a $P(\mathbf{O}|\mathbf{W_i})$, como:

$$p(\mathbf{O}|\lambda_i) = \sum_{\mathbf{X^i}} a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} p(\mathbf{o_t}|x_t^i) a_{x_t^i x_{t+1}^i} \right] \tag{1.8}$$

donde el sumatorio se extiende a todas las posibles secuencias de estados, $\mathbf{X^i}$, que recorren el modelo λ_i; cualquier secuencia de estados está representada por todos los estados internos que la constituyen, $\mathbf{X^i} = \{x_0^i, \ldots, x_T^i\}$; $a_{x_t^i x_{t+1}^i}$ representa la probabilidad de transición del estado x_t^i al x_{t+1}^i; y, por último, $p(\mathbf{o_t}|x_t^i)$ representa la verosimilitud de que el estado x_t^i emita la observación $\mathbf{o_t}$. Esta ecuación lleva implícita la asunción de independencia entre observaciones en distintos instantes de tiempo.

Evaluar todas las secuencias de estados posibles acarrea un coste demasiado alto por lo que se aproxima la probabilidad descrita en la ecuación (1.8) por la evaluación

de únicamente la secuencia de estados más probable:

$$\hat{p}(\mathbf{O}|\lambda_i) = \max_{\mathbf{X^i}} a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} p(\mathbf{o_t}|x_t^i) a_{x_t^i x_{t+1}^i} \right] \tag{1.9}$$

Por otro lado, es habitual trabajar con el logaritmo de esta probabilidad, con lo que la expresión anterior se transforma en la siguiente:

$$\log \hat{p}(\mathbf{O}|\lambda_i) = \max_{\mathbf{X^i}} \left\{ \log(a_{x_0^i x_1^i}) + \sum_{t=1}^{T} \left[\log(p(\mathbf{o_t}|x_t^i)) + \log(a_{x_t^i x_{t+1}^i}) \right] \right\} \tag{1.10}$$

Para poder evaluar esta última expresión necesitamos describir la distribución de probabilidad que gobierna las emisiones en cada estado. Es común que esta distribución se modele mediante una mezcla de Gausianas, que adopta la siguiente expresión:

$$p(\mathbf{o_t}|x_t^i) = p(\mathbf{o_t}|j) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{o}; \mu_{\mathbf{jm}}, \mathbf{\Sigma_{jm}}) \tag{1.11}$$

donde hemos cambiado la notación del estado x_t^i por otra más breve y conveniente, j. En esta ecuación, M indica el número de mezclas; $\mu_{\mathbf{jm}}$ y $\mathbf{\Sigma_{jm}}$ el vector de medias y la matriz de covarianzas que representan a la mezcla m en el estado j. Otra hipótesis habitual es emplear matrices de covarianzas diagonales:

$$p(\mathbf{o_t}|j) = \sum_{m=1}^{M} c_{jm} \prod_{k=1}^{N} \mathcal{N}(o_{t_k}; \mu_{jm_k}, \sigma_{jm_k}^2), \tag{1.12}$$

donde N es la dimensión del vector de parametrización $\mathbf{o_t}$; o_{t_k} su k-ésima componente; μ_{jm_k} representa la k-ésima componente del vector de medias $\mu_{\mathbf{jm}}$ asociado a la mezcla m del estado j; y, por último, $\sigma_{jm_k}^2$ representa la k-ésima componente de la diagonal de la matriz de covarianzas $\mathbf{\Sigma_{jm}}$ asociada a la mezcla m del estado j. Es decir,

$$\mathcal{N}(o_{t_k}; \mu_{jm_k}, \sigma_{jm_k}^2) = \frac{1}{\sqrt{2\pi\sigma_{jm_k}^2}} e^{-\frac{1}{2}\frac{(o_{t_k} - \mu_{jm_k})^2}{\sigma_{jm_k}^2}} \tag{1.13}$$

La fase de entrenamiento de los sistemas basados en HMMs permite, a través de ejemplos, estimar los parámetros que caracterizan a los HMMs. Concretamente, se

determinan las probabilidades de transición entre estados a_{lj} y los parámetros que definen la mezcla de Gausianas (c_{jm}, μ_{jm_k}, $\sigma^2_{jm_k}$). Para esta estimación se emplea el algoritmo de *Baum-Wech* que, de manera iterativa, estima estos parámetros de forma eficiente.

1.1.3. Modelo de Lenguaje

El término $P(\mathbf{W_i})$ que interviene en la ecuación (1.4) recibe la denominación de modelo de lenguaje. Así, a través del modelo de lenguaje obtenemos las probabilidades a priori de las secuencias de palabras a reconocer. Para estimar este valor para secuencias de cualquier longitud necesitaríamos una gran cantidad de datos por lo que debemos acudir a aproximaciones.

Las aproximaciones que probablemente están más extendidas son las basadas en n-gramas. En estos tipos de modelos de lenguaje la probabilidad de aparición de una palabra únicamente depende de un número reducido de palabras que la preceden. En esta tesis únicamente se han usado modelos de lenguaje que tienen en cuenta relaciones entre sólo dos palabras; este caso se corresponde con el empleo de una bigramática ($n = 2$). Así, este tipo de modelos de lenguaje únicamente emplean la palabra que precede a la actual para determinar la probabilidad de la secuencia. De este modo, la probabilidad a priori de la secuencia de palabras $\mathbf{W_i}$ se aproxima mediante la siguiente relación:

$$P(\mathbf{W_i}) \approx P(w_1) \prod_{k=2}^{T} P(w_k|w_{k-1}) \tag{1.14}$$

Los valores que intervienen en la ecuación anterior se estiman usando ejemplos obtenidos a partir de bases de datos de textos.

1.1.4. Decodificación

En la fase de reconocimiento debemos resolver el problema planteado por la ecuación (1.4), es decir, debemos obtener la secuencia de palabras que genera con mayor

verosimilitud la señal de voz a la entrada. Para realizar el reconocimiento nos apoyamos en un sistema basado en HMMs y empleamos los modelos acústicos y de lenguaje que acabamos de describir. De este modo se resuelve el problema alternativo de encontrar el modelo HMM (generalmente construido concatenando varios modelos HMMs menores) con la secuencia de estados que se ajusta de manera más verosímil a las observaciones de entrada. Matemáticamente, este criterio se expresa como sigue:

$$\lambda = \arg\max_i p(\lambda_i|\mathbf{O}) = \arg\max_i p(\lambda_i)\hat{p}(\mathbf{O}|\lambda_i) \tag{1.15}$$

siendo λ el modelo ganador; λ_i el i-ésimo modelo candidato, que recordemos está asociado a la secuencia de palabras $\mathbf{W_i}$; $p(\lambda_i)$ es la probabilidad a priori de esa secuencia de palabras ($P(\mathbf{W_i})$) y viene determinada por el modelo de lenguaje; y, por último, $\hat{p}(\mathbf{O}|\lambda_i)$ representa la probabilidad de que la secuencia de estados más probable del modelo λ_i haya generado la secuencia de observaciones $\mathbf{O}$ (que viene dado por el modelo acústico). Debemos incidir de nuevo en que esta última probabilidad, $\hat{p}(\mathbf{O}|\lambda_i)$, es una aproximación de la probabilidad de que el modelo λ_i haya generado las observaciones $\mathbf{O}$ ya que, para evaluar esta probabilidad, deberíamos considerar todos los caminos posibles y no sólo el más probable.

El algoritmo de *Viterbi* es capaz de resolver esta maximización de forma eficiente. Este algoritmo parte de un estado inicial y, teniendo en cuenta la probabilidad de transición entre estados, la probabilidad de emisión de estos estados y las probabilidades que gobiernan la concatenación de modelos representativos de palabras, obtiene de manera recurrente la secuencia de estados más probable. Los modelos que subyacen a esta secuencia de estados más probable son los que determinan la trascripción de la secuencia que estamos reconociendo. Si denotamos como $\phi_j(t)$ la probabilidad de estar en el estado j y haber observado los t primeros vectores de parámetros recorriendo el camino de máxima verosimilitud, el algoritmo de *Viterbi* puede describirse a través de la siguiente recursión:

$$\phi_j(t) = \max_l\{\phi_l(t-1)a_{lj}\}p(\mathbf{o_t}|j) \tag{1.16}$$

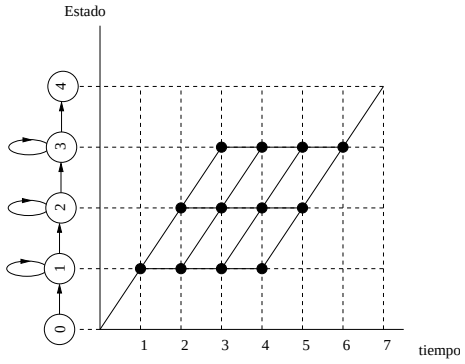


Figura 1.8: Ilustración del algoritmo de *Viterbi*.

donde

$$\phi_1(1) = 1 \tag{1.17}$$

$$\phi_j(1) = a_{1j} p(\mathbf{o_1}|j) \quad \text{para } 1 < j < P \tag{1.18}$$

siendo el estado 1 el estado inicial y P el final. Por tanto, la probabilidad de la secuencia de estados más probable viene dada por:

$$\hat{p}(\mathbf{O}|\lambda_i) = \phi_P(T) = \max_l \{\phi_l(T) a_{lP}\} \tag{1.19}$$

Este algoritmo se puede visualizar por medio de una rejilla como la mostrada en la Figura 1.8. El eje vertical representa a los estados y el eje horizontal el trascurso del tiempo. Esa rejilla muestra todos los caminos posibles que enlazan el nodo inicial con el final. Cuando dos caminos confluyen al mismo nodo únicamente sobrevive el camino que hasta ese punto era el más probable.

Aunque no está explícitamente indicado en las ecuaciones (1.16)-(1.18), en el caso en el que reconocemos habla continua habrá ocasiones en que las transiciones ocurren entre estados finales e iniciales de palabra. En esos casos, para un modelo de lenguaje bigramático, podemos considerar que el término a_{lj} que modela la probabilidad de transición entre estos estados viene dada por la probabilidad de que la palabra representada por el modelo que abandonamos vaya seguida por la palabra representada por el modelo al que vamos.

Es práctica común introducir en el reconocedor constantes de diseño que permiten establecer un compromiso entre la importancia que se asigna al modelo acústico y al modelo de lenguaje. Así, los valores obtenidos a través del modelo de lenguaje que gobernarían las transiciones entre palabras son ponderadas para equilibrar su contribución a la decisión final. Por otro lado, para evitar que las palabras breves pero muy probables tengan una aparición excesiva se añade una penalización cada vez que se inserta una palabra nueva en la secuencia transcrita. Ambas constantes se suelen determinar de forma empírica en los sistemas de reconocimiento de modo que el número de errores por sustitución de palabras esté equilibrado con el número de inserciones que introduce el reconocedor.

1.2. Motivación y objetivos de la tesis

Uno de los problemas más importantes en reconocimiento de habla es el desajuste entre las condiciones acústicas de entrenamiento y de test (funcionamiento). Dada la diversidad de entornos donde actúan los sistemas de reconocimiento de habla, resulta impracticable diseñar un sistema capaz de anticiparse a cualquier condición acústica. En esta tesis proponemos soluciones a esta problemática cuando asumimos que el entorno acústico introduce ruido en la señal de voz y cuando dicho entorno es el propio de un sistema de comunicaciones inalámbrico. De este modo, el objetivo principal de esta tesis consiste en buscar soluciones originales y eficaces al problema de reconocimiento de habla en presencia de ruidos aditivos y de las distorsiones propias de los sistemas de comunicaciones inalámbricos actuales.

Para reducir el efecto de los ruidos aditivos nos inspiramos en una clase de técnicas que basan el reconocimiento en las características de la señal de voz que se consideran más fiables (*missing features*) [Cooke et al., 2001, Raj et al., 2004]. La utilización de estas técnicas lleva a la conclusión de que la robustez de los sistemas de reconocimiento puede aumentarse significativamente si, como paso previo al reconocimiento, desechamos las características de la voz que están más contaminadas. El problema

surge en la detección de estas características corruptas, ya que los errores introducidos en esta detección reducen en gran medida la eficacia de este tipo de técnicas. Sin embargo, en ausencia de errores los resultados son tan prometedores que hace que nos inclinemos por buscar soluciones que palien esta problemática. Adicionalmente, las contribuciones presentadas en esta tesis se basan en la parametrización *Frequency Filtered* [Nadeu et al., 1995, Nadeu et al., 2001] esperando que su sencillez analítica nos permita modelar matemáticamente los efectos de este tipo de distorsiones.

Por otro lado, el problema del reconocimiento de habla con sistemas de comunicación inalámbricos se debe abordar de distinta manera ya que su modelado matemático es mucho más complejo. Así, se estudiará el efecto de estos sistemas en el espectro de modulación y, tras observar este efecto, se propondrán soluciones que eviten la degradación de los sistemas de reconocimiento.

1.3. Estructura de la tesis

Antes de proceder a describir las propuestas realizadas en el marco de esta tesis, en el Capítulo 2 repasamos el estado del arte en reconocimiento robusto de habla. A lo largo de este capítulo se revisan las principales técnicas que han sido propuestas en la literatura con el propósito de conseguir sistemas robustos ante las típicas distorsiones que padece la señal de voz. Esta descripción de técnicas nos servirá como base para explicar nuestras propuestas.

El resto de los capítulos se centran en las aportaciones realizadas en la tesis. En concreto, los Capítulos 3 y 4 tratan el efecto de los ruidos aditivos, mientras que el Capítulo 5 estudia las distorsiones típicas de una comunicación inalámbrica.

El Capítulo 3 se centra en la detección y tratamiento de los *outliers*. En este capítulo partimos de una técnica propuesta por [Matsui and Furui, 1992] en un entorno de reconocimiento de locutor y nosotros la aplicamos a tareas de reconocimiento automático de habla denominándola *bounded distance HMM*. Además, proponemos la combinación de esta técnica con sustracción espectral como medio para superar

algunas de sus limitaciones. Por último, en este capítulo realizamos un análisis detallado de las razones por las que la combinación de estos dos métodos es tan efectiva.

Sustracción espectral deja un cierto nivel de incertidumbre en las observaciones que no se tiene en cuenta en los reconocedores convencionales. En el Capítulo 4 incorporamos esta incertidumbre para el caso particular de la parametrización *Frequency Filtered*. Gracias a esta modificación disminuimos la importancia atribuida a las características de la rejilla tiempo-frecuencia que están más contaminadas. De nuevo, estas técnicas las aplicamos en combinación con *bounded distance HMM*, de modo que también compensamos el efecto de los *outliers*.

En el Capítulo 5 se estudian los efectos de un canal inalámbrico sobre los sistemas de reconocimiento de habla. En este capítulo presentamos una propuesta basada en el filtrado temporal de las características espectrales que representan la señal de voz. Además, el filtrado paso-banda que proponemos también se combina con *bounded distance HMM* para compensar los *outliers* debidos a la presencia de un canal de transmisión ruidoso.

Por último, en el Capítulo 6 presentamos las principales conclusiones alcanzadas a lo largo de esta tesis y las lineas de investigación futuras.

Capítulo 2

Reconocimiento Robusto de Habla

2.1. Introducción

La tasa de acierto en palabras en un sistema de reconocimiento automático de habla (RAH) decae rápidamente cuando la señal de voz sufre distorsiones que no han sido consideradas en la etapa de entrenamiento. [Lippmann, 1997] revisa experimentos que se han realizado para comparar la capacidad de reconocimiento de los seres humanos y de los sistemas de reconocimiento. Tanto en ausencia como en presencia de distorsiones los seres humanos producen resultados significativamente mejores que los sistemas actuales de reconocimiento de habla. Sin embargo, advierte que las diferencias entre un entorno limpio y otro contaminado son especialmente críticas en los sistemas actuales y, mientras que los seres humanos somos capaces de mantener altas tasas de reconocimiento en presencia de ruido, estas tasas decaen rápidamente en los sistemas actuales. Queda claro por tanto que la robustez en los sistemas de reconocimiento debe seguir aumentando.

Muchos han sido los intentos por paliar los efectos del cambio de entorno entre el entrenamiento y el reconocimiento y en este capítulo repasaremos algunas de las principales técnicas propuestas durante los últimos años. El repaso de técnicas no pretende ser exhaustivo, sino tan sólo pretende presentar en primer lugar las grandes

lineas de trabajo en reconocimiento robusto y en segundo lugar, prestar especial atención a aquéllas que están directamente relacionadas con las propuestas realizadas en esta tesis.

La estructura seguida en este capítulo está basada en la clasificación de métodos realizada por [Gong, 1995]. [Gong, 1995] clasifica los métodos que persiguen sistemas robustos en tres categorías: técnicas basadas en parametrizaciones robustas, técnicas basadas en métodos de reducción de ruido y, finalmente, técnicas que adaptan los modelos del reconocedor al nuevo entorno de aplicación. Como veremos a lo largo del capítulo esta clasificación no es cerrada y, en ocasiones, encontraremos métodos que bien podrían pertenecer a varias categorías de forma simultánea. Adicionalmente, hemos estudiado dos grupos de técnicas que, si bien podrían encajar en alguna de las categorías propuestas por [Gong, 1995], preferimos estudiarlas de forma independiente debido a los estrechos vínculos que poseen con las propuestas realizadas en esta tesis. Estos métodos se corresponden con los que basan el reconocimiento en las características del espectrograma más fiables y los que incorporan la incertidumbre de los parámetros en el proceso de reconocimiento.

2.2. Parametrizaciones robustas

Englobamos en esta categoría todas las técnicas que buscan un conjunto de parámetros lo más invariante posible ante distorsiones tales como el ruido aditivo o el cambio de micrófono.

A su vez, dividimos esta categoría en otras basándonos en la forma de abordar el problema. En un primer lugar estudiamos técnicas que normalizan los parámetros de entrada al reconocedor de modo que presenten los mismos momentos estadísticos independientemente del entorno de aplicación. En segundo lugar, describimos técnicas que seleccionan las componentes del espectro de modulación más robustas y, de este modo, eliminan aquéllas que poseen información irrelevante para la tarea de reconocimiento. En tercer lugar, revisamos algunas técnicas que modifican la medida

de similitud entre modelos y observaciones, de modo que el reconocimiento se basa en los coeficientes de los vectores de parametrización que permanecen más estables. Finalmente, presentamos el algoritmo de Viterbi ponderado que, aunque no es estrictamente una parametrización robusta, concede más importancia a las tramas que permanecen, en mayor medida, inalteradas.

2.2.1. Normalización de los parámetros

Englobamos dentro de esta categoría a aquellas técnicas que normalizan los descriptores estadísticos de los parámetros de entrada. Dentro de este grupo, CMN (*Cepstral Mean Normalization* [Furui, 1981]) es probablemente la técnica más conocida y extendida. CMN se basa en el hecho de que una distorsión convolutiva en el dominio del espectro se transforma en aditiva en el dominio del cepstrum. Este tipo de distorsiones convolutivas son típicas de los canales de transmisión (ej. micrófonos) y se pueden compensar imponiendo que los parámetros a la entrada del reconocedor presenten media nula. Así, la normalización propuesta por CMN para los parámetros es

$$\hat{o}_k = o_k - \mu_{o_k} \tag{2.1}$$

donde $\hat{o}_k$ representa a la componente k-ésima del vector de observaciones tras aplicar CMN, o_k dicha componente antes de la normalización y μ_{o_k} su media. Normalmente esta media se calcula para cada frase.

A diferencia de CMN, [Viikki and Laurila, 1998] proponen normalizar los momentos de primer y segundo orden, es decir, la media y la varianza de los parámetros. Así, esta técnica, conocida como MVN (*Mean and Variance Normalization*), aumenta la robustez de los sistemas frente a ruidos convolutivos y aditivos. La expresión que rige esta normalización es la siguiente:

$$\hat{o}_k = \frac{o_k - \mu_{o_k}}{\sigma_{o_k}} \tag{2.2}$$

donde μ_{o_k} es de nuevo la media de los parámetros y $\sigma^2_{o_k}$ su varianza. Es común calcular

esta media y esta varianza enventanando la trayectoria temporal de las observaciones con muestras pasadas y futuras de la observación actual.

Otros autores van más allá y normalizan todos los momentos de las distribuciones de los parámetros. Así, [de la Torre et al., 2005] proponen el método conocido como HEQ (*Histogram equalization*) que transforma el histograma de los parámetros de entrada al correspondiente a una Gausiana con media 0 y varianza 1. Tanto CMN como MVN pueden únicamente compensar distorsiones lineales en los parámetros, sin embargo HEQ, al compensar todos los momentos, podría compensar distorsiones no lineales lo que la convierte en una técnica más potente. Por último, en [Molau et al., 2003a] se propone un planteamiento similar basado en la normalización de histograma, de nuevo este trabajo concluye que la normalización de histograma es más potente que la normalización de los dos primeros momentos de la distribución de los parámetros. Además en [Molau et al., 2003b] estudian la etapa más idónea, dentro de las diferentes etapas involucradas en la obtención de los parámetros cepstrales, para realizar esta normalización. Concluyen que es preferible realizar la normalización en el dominio de las log-energías en banda frente a realizarla en la última etapa, donde se calculan los parámetros cepstrales. Por otro lado, destacan la importancia de proponer funciones de normalización que tengan en cuenta la proporción de silencio presente en las locuciones.

2.2.2. Filtrado del espectro de modulación

Otro grupo de técnicas que buscan parametrizaciones robustas ante distorsiones son las que filtran el espectro de modulación. La Figura 2.1 representa de forma gráfica el concepto de espectro de modulación. Como vemos en dicha figura, este espectro es el de la señal generada por la evolución temporal de cada coeficiente de nuestra parametrización.

Las técnicas basadas en el filtrado del espectro de modulación persiguen preservar todas aquellas frecuencias de modulación con información lingüística relevante, mientras que deben rechazar aquellas con información no relevante para el recono-

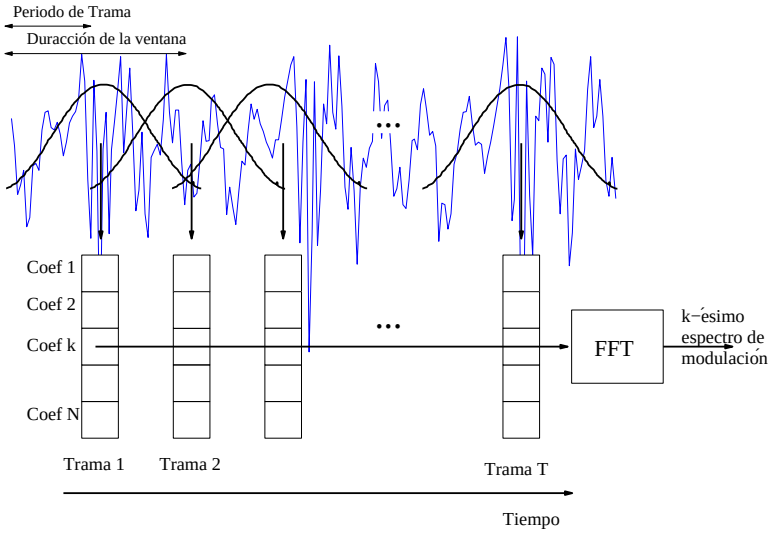


Figura 2.1: Espectro de modulación para el k-ésimo coeficiente

cedor (p. ej. componentes con información sobre el canal de comunicación o con información sobre el ruido). Esta idea está directamente relacionada con varios experimentos perceptuales en los que se demuestra que las componentes que aportan inteligibilidad a la señal de habla están concentradas en torno a ciertas bandas del espectro de modulación mientras que el resto no contribuye de forma significativa [Greenberg, 1996]. Típicamente la supresión de las componentes menos importantes del espectro de modulación se lleva a cabo mediante un filtrado de las trayectorias temporales de los parámetros.

La técnica CMN [Furui, 1981], explicada en la Sección 2.2.1, puede interpretarse como una de las primeras técnicas en abordar este concepto. CMN suprime la frecuencia de modulación correspondiente a los 0 Hz y así aumenta la robustez del sistemas reconocedor, especialmente ante distorsiones convolutivas.

Seguramente sea RASTA (RelAtive SpecTrAl [Hermansky and Morgan, 1994]) el primero de los métodos que hace un énfasis explícito en favorecer las frecuencias más significativas del espectro de modulación. Básicamente, RASTA realiza un filtrado paso banda de las log-energías en bandas de modo que preserva las frecuencias de modulación comprendidas entre 1 y 12 Hz. La porción paso bajo del filtro ayuda

a suavizar los cambios abruptos en el espectro debidos al propio análisis llevado a cabo para extraer los parámetros. La parte paso alto se diseña inicialmente para minimizar el efecto de los ruidos convolucionales, tales como el cambio de micrófono o del sistema de transmisión (RASTA se propuso para telefonía). Este ruido convolutivo se transforma en aditivo en el dominio del log-espectro. Además, su rango de variación se sitúa a frecuencias de modulación distintas de las que ocupa el espectro de la señal de voz y, por tanto, son eliminadas mediante RASTA. De hecho, [Hermansky and Morgan, 1994] demuestran que al reducir esta información se mejoran de forma significativa los resultados en el reconocedor.

[Hanson and Applebaum, 1993] proponen utilizar un filtrado paso banda tipo RASTA y otro paso alto. De nuevo, estudian distorsiones producidas por el canal, ruidos aditivos y debidas al conocido efecto Lombard. Al estudiar ambos tipos de filtros llegan a la conclusión de que es la parte paso alto la responsable de las mejoras en el sistema de reconocimiento. Además, los filtros son aplicados tanto a las log-energías en banda como a los coeficientes cepstrales encontrando resultados equivalentes: esto se debe a que los coeficientes cepstrales se obtienen aplicando una Transformada Discreta de Coseno (DCT), que es un operador lineal, a las log-energías en banda. Esto significa que se pueden aplicar técnicas de filtrado sobre parametrizaciones que no calculan las energías en banda, como pueden ser las parametrizaciones basadas en los coeficientes LPC [Smolders and Van Compernolle, 1993].

Otros autores proponen filtros más sofisticados. Por ejemplo, [Nadeu et al., 1997] emplean una cascada de filtros compuesta por un filtro igualador de primer orden y un filtro paso banda; dicha cascada de filtros se aplica a la parametrización LPC-cepstrum. Los autores concluyen que favorecer frecuencias alrededor de los 3 Hz es beneficioso para el sistema de reconocimiento de habla. Es interesante destacar que esos 3 Hz coinciden con la tasa media de sílabas en la base de datos de sus experimentos.

[Kanedera et al., 1998] realizan un estudio sobre la relevancia de las bandas en el espectro de modulación. Las principales conclusiones que extraen son las siguientes:

- En entornos limpios, la mayor parte de la información está contenida entre 1 Hz y 16 Hz del espectro de modulación.

- La banda alrededor de los 4 Hz resulta ser la componente de mayor utilidad tanto con voz limpia como en presencia de alguna distorsión (resultado similar a [Nadeu et al., 1997]).

- En un ambiente contaminado, las componentes del espectro de modulación por debajo de 2 Hz y por encima de 10 Hz son poco importantes para la inteligibilidad de la señal de habla. En particular, las componentes por debajo de 1 Hz contienen sobre todo información del entorno (ej. efectos del canal de transmisión). Por tanto, suprimiendo estas frecuencias se pueden obtener mejores tasas de reconocimiento.

Algunos autores [Hanson and Applebaum, 1993, Nadeu et al., 2001] han hecho hincapié en la relación entre las técnicas basadas en el filtrado del espectro de modulación y los coeficientes dinámicos de primer y segundo orden [Furui, 1986]. De hecho, estas características pueden interpretarse como el resultado de aplicar un filtro paso alto a las características estáticas cepstrales para favorecer las frecuencias alrededor de los 10 Hz. Esté nuevo punto de vista explicaría su efectividad ante distorsiones tanto convolutivas como aditivas.

2.2.3. Ponderación de las medidas de similitud entre vectores de características

De entre los coeficientes que forman el vector de parametrización hay unos que son más robustos que otros. Por ello, encontramos un nuevo grupo de técnicas dentro de esta categoría que basan el reconocimiento en aquellos coeficientes menos alterados por las distorsiones que afectan al reconocimiento. Para ello, modifican la medida de similitud empleada para comparar las observaciones actuales con las que caracterizan los modelos en el reconocedor. [Soong and Sondhi, 1988] explotan este concepto

para mejorar las prestaciones de un reconocedor basado en DTW (*Dynamic Time Warping*). Así, modifican la distancia de Itakura para dar más importancia a los picos del espectro, más robustos ante el ruido, que a los valles. De este modo, esta nueva distancia puede expresarse como:

$$d = \log \int_{-\pi}^{\pi} F(\omega) \frac{|B(\omega)|^2}{|A(\omega)|^2} \frac{d\omega}{2\pi} \tag{2.3}$$

donde $\frac{1}{B(\omega)}$ es el espectro LP normalizado en energía tomado como referencia y $\frac{1}{A(\omega)}$ el espectro LP, también normalizado, que estamos clasificando. $F(\omega)$ es una función de ponderación que está directamente relacionada con el espectro LP y que, por tanto, pondera en mayor medida los picos del espectro respecto a los valles. Aunque los resultados obtenidos empleando esta medida muestran la importancia de introducir esta ponderación, esta nueva distancia es únicamente válida para comparar espectros LP y, si se emplean otro tipo de parametrizaciones, debemos acudir a otro tipo de ponderaciones.

Así, se proponen otras medidas para aumentar la capacidad de discriminación de los sistemas de reconocimiento cuando se emplean parámetros cepstrales. Tanto WCP (*Weighted Cepstral Distance* [Tohkura, 1987]) como QWC (*Quefrency-Weighted Cepstral* [Paliwal, 1982]) son dos ejemplos donde la medida de distancia se mejora ponderando los coeficientes. Ambas medidas se diseñaron para funcionar con sistemas de reconocimientos basados en DTW y reducen la importancia de los coeficientes cepstrales bajos frente a los altos. Esto se debe a que los coeficientes bajos son los que presentan mayor varianza y, por tanto, acaparan un porcentaje demasiado elevado de la capacidad de discriminación del sistema. Por tanto, estos autores proponen emplear una distancia euclídea ponderada:

$$d = \sum_{k=1}^{D} w_k \left(o_k^{test} - o_k^{ref} \right)^2 \tag{2.4}$$

donde o_k^{ref} y o_k^{test} son, respectivamente, los k-ésimos coeficientes cepstrales de las tramas de referencia y a reconocer; w_k representa la función de ponderación que crece con el índice cepstral.

Por otro lado, en [Juang et al., 1987] encontramos razones para elegir otro tipo de ponderaciones en presencia de distorsiones. Así, el estudio realizado por [Juang et al., 1987] evalúa distintos tipos de liftering y los analiza teniendo en cuenta su distancia equivalente en el sistema de reconocimiento. Llega a la conclusión de que ni WCP ni QWC son medidas adecuadas ya que los coeficientes cepstrales de órdenes altos se ven más afectados por el ruido y estas medidas enfatizan estos coeficientes. Además, los cambios del sistema de transmisión afectan sobre todo a los coeficientes bajos, por lo que propone un liftering paso-banda para compensar todos estos efectos.

Las mismas ideas basadas en dar mayor importancia a los coeficientes más robustos han sido también aplicadas a sistemas basados en HMMs (*Hidden Markov Models*). [Carlson and Clements, 1991] proponen un método que compensa la reducción de la norma de los parámetros cepstrales en presencia de ruido blanco. Este método proyecta los vectores de entrada contaminados al espacio de los vectores representativos de los modelos HMM. Así, modifican la distancia Euclídea que encontramos en el exponente de las distribuciones Gausianas de los modelos por otra que mide el error de reconstrucción cometido al realizar esta proyección. Finalmente, esto equivale a modificar el vector de medias de los modelos mediante un factor que cuantifica dicha proyección. Por otro lado, se observa que esta medida enfatiza las zonas de alta energía del espectro por lo que, de nuevo, encontramos un sistema que se apoya en las zonas del espectro con mejor relación SNR. Además, esta proyección da más importancia a las tramas con mayor energía frente a las que tienen una energía menor. Sin embargo, esta nueva técnica presenta el problema de tender a distancias próximas a cero cuando los vectores de entrada están muy contaminados. [Chien et al., 1995] corrigen este defecto modificando también la varianza de los modelos. De este modo, tanto las medias como las varianzas de los modelos son modificadas para tener en cuenta el ruido presente en los vectores a reconocer.

2.2.4. Algoritmo de Viterbi ponderado (Weighted Viterbi)

Los métodos basados en el algoritmo de Viterbi ponderado (*Weighted Viterbi*) no se encuadran estrictamente en el ámbito de las parametrizaciones robustas, pero lo incluimos aquí porque favorecen o penalizan unos vectores de parametrización frente a otros, lo que resulta en consonancia con lo perseguido por los métodos explicados en las secciones anteriores. Así, estos métodos dan más importancia a las tramas más robustas modificando el algoritmo de Viterbi del siguiente modo:

$$\phi_j(t) = \max_l\{\phi_l(t-1)a_{lj}\}(p(\mathbf{o}_t|j))^{\gamma_t} \tag{2.5}$$

donde $\phi_j(t)$ representa la máxima verosimilitud de observar el vector $\mathbf{o_t}$ y estar en el estado j en el instante de tiempo t; a_{lj} es la probabilidad de transición entre el estado l y el j; $p(\mathbf{o}_t|j)$ la probabilidad de emitir $\mathbf{o}_t$ en el estado j y, por último, γ_t es el factor de ponderación que introduce el algoritmo de Viterbi ponderado para aportar o restar énfasis a las tramas temporales.

El factor γ_t ha adoptado distintas expresiones. [Yoma et al., 1995] asignan a este parámetro un valor que depende de la SNR de la trama actual y, en otros experimentos, otro que depende de la periodicidad medida en dicha trama. De este modo, consideran más robustas, por un lado, las tramas con mayor SNR y, por otro, las tramas sonoras. [Yoma et al., 1998] cambian este factor de ponderación por otro inversamente proporcional a la incertidumbre existente en las observaciones tras aplicar sustracción espectral. [Cui and Alwan, 2004], tras una estimación de los parámetros, emplean como medida de ponderación la distancia entre la media de los parámetros a la entrada al reconocedor y la media almacenada en los modelos. De este modo, si la estimación no fue precisa habrá una gran diferencia en las medias, lo que sería indicativo de la baja fiabilidad de las observaciones.

No sólo ante ruidos aditivos se ha adoptado esta metodología, [Bernard and Alwan, 2002] emplean el Viterbi ponderado con el objetivo de reducir el impacto de los errores de transmisión y de la pérdida de tramas en los sistemas de reconocimiento de habla. Los experimentos se realizan en un entorno

distribuido: el terminal parametriza la señal de voz y transmite los vectores de parámetros mientras que, el servidor, recibe estos vectores y aborda el proceso de reconocimiento. Como medida de lo distorsionados que están los vectores recibidos proponen dos: la primera mide la distancia entre el vector recibido y los dos vectores-código más cercanos de modo que, si estas distancias son parejas, el vector actual es poco fiable; la segunda medida tiene en cuenta el número de tramas consecutivas descartadas por el decodificador, considerando menos fiables los vectores de entrada si el contador de tramas descartadas ha aumentado. Ambas ponderaciones conducen a sistemas más robustos ante las distorsiones introducidas por los sistemas de transmisión.

2.3. Regeneración de parámetros (feature enhancement)

Englobamos en esta categoría las técnicas que estiman los parámetros limpios a partir de los contaminados antes de abordar el reconocimiento.

En esta sección empezamos describiendo la técnica de sustracción espectral, que asume que el ruido es aditivo. Hecha esta suposición, es posible compensar los parámetros contaminados en el dominio del espectro ya que en este dominio la distorsión también será aditiva. A continuación, describimos técnicas que no se limitan a ruidos aditivos y que compensan directamente el vector de características que, típicamente, se corresponde con un vector de coeficientes cepstrales. Estas técnicas tienen una complejidad mayor que sustracción espectral ya que no hay una relación sencilla entre los parámetros contaminados y los limpios.

2.3.1. Estimación en el dominio del espectro: Sustracción espectral

La sustracción espectral es probablemente la técnica más clásica y conocida dentro de esta categoría. Esta técnica asume que la señal de voz está contaminada mediante un ruido aditivo incorrelacionado con la misma. Así, estima el espectro de la voz limpia sustrayendo al de la voz contaminada una estimación del espectro del ruido. En la literatura encontramos diversas versiones que se basan en los mismos principios pero que presentan ligeras diferencias. En el artículo escrito por [Gong, 1995] encontramos un interesante resumen con las principales propuestas. Aquí explicamos brevemente algunas de las diferencias destacables entre unas y otras:

- Cuando se emplea sustracción espectral para mejorar la robustez de los sistemas de reconocimiento de habla, en lugar de estimar el espectro de la voz limpia, se estima o bien su módulo o bien su densidad espectral de potencia. Esto se debe a que la fase del espectro no lleva información útil para discriminar los sonidos y, por tanto, no se utiliza en los sistemas de reconocimiento de habla.

- Cuando la estimación del espectro del ruido o, más concretamente, de su módulo o densidad espectral de potencia, se sustrae del de la voz contaminada pueden producirse valores negativos. Lógicamente, ni el módulo ni la potencia del espectro pueden tomar valores negativos con lo que se imponen correcciones. Pese a que encontramos diferentes alternativas en la literatura, es común imponer un valor mínimo al espectro de la señal proporcional a la estimación del espectro de ruido (el factor de proporcionalidad se denomina "nivel de suelo" (*spectrum flooring*)). Esta saturación introduce efectos no lineales en el espectro reconstruido que reducen la calidad de la señal de voz estimada; típicamente, se aprecian ruidos poco naturales conocidos con el nombre de "ruido musical".

- Con el objetivo de reducir el "ruido musical" y el ruido remanente en la señal de voz, [Berouti et al., 1979] proponen sustraer, en lugar de la estimación del

espectro del ruido, esta estimación pero ponderada por un factor mayor que 1 ("factor de sobre-estimación", *over-estimation factor*). Por un lado, esta sustracción tan agresiva reduce el ruido remanente y, por otro lado, imponiendo un valor mínimo al espectro tal y como vimos en el punto anterior, evitamos que el espectro reconstruido presente picos estrechos pero de potencia elevada tan característicos del "ruido musical".

- Una buena estimación del espectro del ruido es determinante para asegurar el éxito de los métodos basados en sustracción espectral. Encontramos dos vertientes principales en la estimación del espectro del ruido. Por un lado, las que confían en un detector de actividad vocal (VAD, *Voice Activity Detector*), que estiman el ruido en los periodos de silencio presentes en la señal de voz; estas técnicas asumen que el ruido varía más lentamente que la señal de voz de modo que las estimación realizadas durante el silencio representan al ruido durante los periodos de actividad vocal; estas técnicas, cómo es lógico, además dependen de la efectividad del VAD que clasifica las tramas como silencio o voz. Por otro lado, encontramos métodos que no se basan en este tipo de clasificadores (p. ej. [Martin, 2001, Stahl et al., 2000]) y suelen tener en cuenta que la señal de voz, incluso en periodos de actividad, suele decaer a niveles representativos del ruido.

De este modo, un típico algoritmo de sustracción espectral que actúa sobre la densidad espectral de potencia pasa a formularse como:

$$\widehat{P_X}(\omega, l) = \max\left\{\widehat{P_S}(\omega, l) - \gamma\widehat{P_N}(\omega, l), \beta\widehat{P_N}(\omega, l)\right\} \tag{2.6}$$

donde $\widehat{P_X}(\omega, l)$ representa la estimación de la potencia del espectro de la voz limpia a la frecuencia ω y el instante (o trama) l; $\widehat{P_S}$ representa la potencia del espectro contaminado; $\widehat{P_N}$ es la estimación de la potencia del espectro de ruido; y finalmente, γ y β son constantes de diseño que fueron introducidas con anterioridad como "factor de sobre-estimación" (*over-estimation factor*) y "nivel de suelo" (*spectrum flooring*),

respectivamente.

A pesar de que sustracción espectral es capaz de mejorar la SNR de la señal de voz, este aumento no siempre se traduce en sistemas de reconocimiento más robustos. Esto se debe a que las no linealidades del método introducen distorsiones que degradan las prestaciones de los reconocedores. Para evitar estas distorsiones, [Nolazco Flores and Young, 1994] adaptan los modelos del reconocedor empleando PMC (*Parallel Model Combination*) consiguiendo un buen modelado de esas distorsiones. Otros autores [Shozakai et al., 1997] abogan por entrenar los modelos empleando secuencias que han sido procesadas usando sustracción espectral independientemente de que estas secuencias estén o no contaminadas.

2.3.2. Estimaciones en el dominio cepstral

A diferencia de sustracción espectral existen técnicas que directamente trabajan en el dominio definido por la parametrización. En [Stern et al., 1996] o [Stern et al., 1997] encontramos un compendio de técnicas para compensar distorsiones aditivas y convolutivas en el dominio cepstral. Entre las técnicas allí descritas, CDCN (*Codeword-Dependent Cepstral Normalization* [Acero and Stern, 1990]) y VTS (*Vector Taylor Series* [Moreno et al., 1996]) son quizá las más populares. Los dos métodos estiman los parámetros limpios a partir de los contaminados mediante estimadores de mínimo error cuadrático medio (MMSE, *Minimum Mean Squared Error*). Para ello modelan la distribución de probabilidad de los parámetros de voz limpia mediante una mezcla de Gausianas e imponen un modelo para las distorsiones. La diferencia entre los métodos es que CDCN no tiene en cuenta la varianza del ruido mientras que VTS sí la tiene en cuenta. Para tenerla en cuenta, VTS aproxima la distorsión en el dominio cepstral mediante series de Taylor.

Más métodos han sido propuestos para paliar las distorsiones que afectan a los parámetros, siendo los coeficientes cepstrales los que probablemente han demandado mayor atención. En general, todos los métodos se enfrentan a relaciones no lineales cuando estudian el efecto de las distorsiones en el dominio de las parametrizaciones.

Esto hace que no se encuentren soluciones analíticas cerradas.

Por otro lado, es común emplear estimadores Bayesianos para la estimación de los parámetros limpios. Para ello, los parámetros de la voz limpia deben ser modelados por alguna distribución de probabilidad, siendo común el empleo de mezclas de Gausianas (p. ej. [Stouten et al., 2006, Droppo et al., 2002]). Normalmente, las diferencias entre unos y otros métodos estriban en el modelado de la relación de los parámetros limpios con los contaminados y en la forma de evitar las no linealidades presentes en las ecuaciones.

2.4. Adaptación de modelos

Este conjunto de técnicas adaptan los parámetros que caracterizan los modelos del reconocedor al nuevo ambiente de aplicación. En esta sección explicaremos brevemente dos métodos (MAP y MLLR) que adaptan los parámetros del reconocedor sin asumir ningún modelo de distorsión y otro que asume la presencia de ruido aditivo (PMC).

2.4.1. Técnicas genéricas de adaptación de modelos: MAP y MLLR

[Gauvain and Lee, 1994] adaptan los modelos del reconocedor a nuevos entornos aplicando estimadores que maximizan la probabilidad a posteriori, conocidos como estimadores MAP (*maximum a posteriory*). Buscan modelos que representen de forma más precisa las observaciones que caracterizan el nuevo entorno; para ello, además de apoyarse en estas observaciones, cuyo número suele ser reducido, se apoyan en el conocimiento a priori de los parámetros que caracterizan los modelos. Matemáticamente podemos escribir este criterio como:

$$\theta_{MAP} = \arg\max_{\theta} p(\mathbf{O}|\theta)g(\theta) \tag{2.7}$$

donde $g(\theta)$ representa la función de densidad de probabilidad a priori de los parámetros θ que representan nuestro modelo y $p(\mathbf{O}|\theta)$ la verosimilitud de las observaciones reales, $\mathbf{O}$, dados esos parámetros θ. Cuando deseamos adaptar un sistema de reconocimiento a un nuevo entorno se emplean, como conocimiento a priori, funciones de densidad de probabilidad descritas mediante los parámetros que caracterizan el sistema antes de la adaptación. Conocida $g(\theta)$, se resuelve la ecuación (2.7) y se obtienen los nuevos parámetros que caracterizan cada modelo. La expresión que se obtiene para la media es representativa del concepto de adaptación y clarifica el comportamiento de este método, por lo que la replicamos a continuación:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{jm}} = \frac{\tau_{jm}\boldsymbol{\mu}_{\boldsymbol{jm}} + \sum_{t=1}^{T} z_{jmt}\mathbf{o_t}}{\tau_{jm} + \sum_{t=1}^{T} z_{jmt}} \tag{2.8}$$

donde $\boldsymbol{\mu}_{\boldsymbol{jm}}$ y $\hat{\boldsymbol{\mu}}_{\boldsymbol{jm}}$ son, respectivamente, el vector de medias de la mezcla m en el estado j antes y después de la adaptación; z_{jmt} es la probabilidad de que el vector de observaciones $\mathbf{o_t}$ sea generado por la mezcla m del estado j; T es el número de observaciones y, finalmente, τ_{jm} es una variable de diseño que pondera la probabilidad a priori considerada. Es fácil observar que la ecuación (2.8) no es más que una suma ponderada del conocimiento a priori, $\boldsymbol{\mu}_{\boldsymbol{jm}}$, y la media de las observaciones empleadas en la adaptación, $\sum_{t=1}^{T} z_{jmt}\mathbf{o_t}$.

MAP adapta los modelos del reconocedor de forma individual. De este modo, MAP será más preciso en tanto en cuanto los datos de adaptación sean representativos de todos los modelos dentro del reconocedor. MLLR (*Maximum Likelihood Linear Regression* [Gales and Young, 1996, Gales, 1998]), a diferencia de MAP, propone transformaciones comunes a los parámetros de todos los modelos y, de este modo, se relaja la necesidad de tener datos de adaptación que representen a todos los modelos. Proponen transformaciones lineales tanto de la media como de la varianza con el objetivo de adaptarse a los nuevos entornos a través de estas transformaciones. Así, la expresión para la media adaptada toma la siguiente expresión:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{jm}} = \mathbf{A}\boldsymbol{\mu}_{\boldsymbol{jm}} + \mathbf{b}. \tag{2.9}$$

donde, al igual que antes, $\boldsymbol{\mu_{jm}}$ y $\hat{\boldsymbol{\mu}}_{\boldsymbol{jm}}$ son, respectivamente, los vectores de medias antes y después de la adaptación y, la matriz $\mathbf{A}$ y el vector $\mathbf{b}$ definen la transformación lineal. La ventaja con respecto a MAP es que esas transformaciones se comparten entre varios modelos de modo que necesitamos estimar un menor número de parámetros. Es por ello que MLLR suele ser más adecuada cuando los datos de adaptación son escasos.

2.4.2. Adaptación de los modelos ante distorsiones aditivas: PMC

Ni MAP ni MLLR modelan la distorsión introducida en los parámetros y, por tanto, son métodos de adaptación que podemos considerar genéricos. Es decir, tanto uno como otro se emplean para adaptar el sistema a un nuevo locutor, un nuevo canal de transmisión o, incluso, a un nuevo ambiente. Existen otros métodos menos genéricos que se diseñan para adaptar los sistemas a distorsiones más específicas. Este es el caso de la técnica PMC (*Parallel Model Combination* [Gales and Young, 1996]) que asume una distorsión aditiva y emplea esa hipótesis para actualizar los modelos. La técnica PMC, además de los modelos de la voz limpia sin adaptar, necesita un modelo que caracterice el ruido. Su objetivo final es combinar estos modelos para generar otros que representen de forma más fiel la voz contaminada.

Esta técnica se desarrolló para sistemas basados en la parametrización cepstrum, por ello, antes de combinar los modelos los transforma al dominio del log-espectro donde el efecto del ruido aditivo es más sencillo de caracterizar. A pesar de esta aparente sencillez, las expresiones que relacionan los parámetros contaminados y los limpios son suficientemente complicadas como para tener que acudir a aproximaciones. Una de esas aproximaciones se denomina *Data Driven* PMC. Esta técnica estima los parámetros de los modelos contaminados a partir de observaciones ruidosas generadas de forma artificial; para ello, emplea los modelos de la voz limpia y del ruido, genera muestras representativas de cada uno de ellos y las combina siguien-

do la caracterización matemática de la distorsión aditiva. A continuación, calcula la media y la varianza de estas muestras artificiales, que pasan a ser los parámetros característicos de los modelos que representan la voz contaminada.

2.5. Métodos de reconocimiento basados en las características más fiables (Missing-Features).

En esta sección introducimos otro importante grupo de técnicas basado en detectar las características no fiables en tiempo y frecuencia y evitar que participen en el reconocimiento. Por consiguiente, estas técnicas presentan una estrecha relación con las estudiadas en la Sección 2.2, que fundamentan el reconocimiento en los parámetros que permanecen más invariantes ante las distorsiones que sufre la señal de voz. Las técnicas que presentamos aquí se centran en los parámetros fiables que, lógicamente, serán los que tuvieron un comportamiento más robusto ante tales distorsiones. En el resto de esta sección explicamos brevemente los principales métodos desarrollados en el marco de esta filosofía.

Estas técnicas se basan en la redundancia que existe en la señal de voz y en el hecho de que, debido a esta redundancia, los seres humanos somos capaces de entender el mensaje contenido en la señal de voz incluso cuando se han perdido componentes del espectrograma. Inspirados en este hecho, las técnicas basadas en las características más fiables (*missing-features*) abordan el reconocimiento empleando únicamente las características del espectrograma que se consideran fiables y prescinden del resto.

Según la metodología empleada para eliminar la influencia de las características consideradas como no fiables encontramos dos tipo de aproximaciones [Raj and Stern, 2005]:

- Métodos que modifican el decodificador [Cooke et al., 2001]: las características no fiables son eliminadas de los vectores de entrada y el reconocimiento se aborda con un conjunto incompleto de características. La desventaja de estas

técnicas es que deben usar características en el dominio del espectro, ya que es en ese dominio en el que se realiza la clasificación de muestras fiables y no fiables. Típicamente, estos métodos emplean las log-energías en banda como parametrización, lo que limita las prestaciones de los reconocedores. Dentro de este grupo encontramos dos aproximaciones principales:

- Basadas en la estimación de las características no fiables (*data imputation*): estas técnicas estiman los coeficientes no fiables de cada vector de parametrización usando las descripciones probabilísticas correspondientes a cada estado y los coeficientes fiables. Así, la estimación de las características no fiables se realiza como el valor esperado de observar los coeficientes no fiables dado que conocemos los fiables, esto es:

$$\hat{\mathbf{o}}_{\mathbf{t}}^{\mathbf{nf}} = E_{\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}}|\mathbf{o}_{\mathbf{t}}^{\mathbf{f}},x_t^i}\{\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}}\} = \int p(\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}}|\mathbf{o}_{\mathbf{t}}^{\mathbf{f}}, x_t^i) d\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}} \tag{2.10}$$

  donde $\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}}$ y $\mathbf{o}_{\mathbf{t}}^{\mathbf{f}}$ representan, respectivamente, los coeficientes no fiables y fiables del vector de parametrización observado en el instante t; $p(\mathbf{o}_{\mathbf{t}}^{\mathbf{nf}}|\mathbf{o}_{\mathbf{t}}^{\mathbf{f}}, x_t^i)$ es la función de densidad de probabilidad condicional que nos permite calcular la probabilidad de una componente no fiable conocidos el estado actual (x_t^i) del modelo λ_i y las componentes fiables. Nótese que esta estimación se lleva a cabo dentro del reconocedor por lo que es necesario modificar el algoritmo de decodificación convencional.

  Estas expresiones suponen que no tenemos conocimiento alguno de los valores que toman las características no fiables; si supiésemos, por ejemplo, que estas características deben estar acotadas entre unos valores máximos y mínimos podemos usar esta información para mejorar la estimación. En este caso tomaríamos el valor estimado mediante la ecuación (2.10) siempre que estemos en el rango conocido y saturaríamos la estimación en cualquier otro caso.

- Marginalización (*marginalisation*): el otro método que modifica el reconocedor para desechar las características no fiables se basa en las funciones

de densidad de probabilidad marginales de las características fiables. De este modo, se sustituyen las funciones de densidad de probabilidad que modelan los estados del reconocedor por esas funciones de densidad de probabilidad marginales. Por tanto, la nueva función de densidad de probabilidad que modela los estados viene dada por la siguiente expresión:

$$p(\mathbf{o_t^f}|x_t^i) = \int_{\mathbf{o_t^{nf}}} p(\mathbf{o_t^{nf}}, \mathbf{o_t^f}|x_t^i) d\mathbf{o_t^{nf}} \tag{2.11}$$

Al igual que sucede con el método anterior, en caso de conocer los valores límites para las características no fiables podemos incluir esta información en la ecuación anterior; en tal caso, la integral se evaluaría únicamente en los valores permitidos para los coeficientes no fiables.

- Métodos de estimación de los parámetros de entrada [Raj et al., 2004]: estos métodos primero determinan las regiones en la rejilla de tiempo y frecuencia del espectrograma que son consideradas no fiables. A continuación, estiman estas características no fiables obteniendo un espectrograma reconstruido completo. A partir de este espectrograma completo se puede realizar cualquier transformación que nos lleve a parametrizaciones más adecuadas para abordar el reconocimiento de habla como, por ejemplo, la parametrización MFCC. Debido a este hecho, estos métodos consiguen prestaciones superiores a aquellos que modifican el reconocedor. Dentro de este grupo encontramos dos aproximaciones principales:

  - Reconstrucción basada en correlaciones (*Correlation-based reconstruction*): se asume que el espectrograma de la señal de voz es una realización de un proceso Gausiano estacionario en sentido amplio. Esto nos permite establecer relaciones entre distintos puntos de la rejilla tiempo-frecuencia en el espectrograma. De este modo, una vez caracterizado estadísticamente este proceso estacionario mediante una etapa de entrenamiento, somos capaces de estimar los puntos del espectrograma no fiables basándonos

en los valores de los puntos fiables que se sitúan a su alrededor. Así, la estimación de las características no fiables adopta la siguiente expresión:

$$\hat{\mathbf{o}}^{\mathbf{nf}} = \boldsymbol{\mu}^{\boldsymbol{nf}} + \boldsymbol{\Sigma}_{\mathbf{nf-f}}\boldsymbol{\Sigma}_{\mathbf{f-f}}^{-\mathbf{1}}(\mathbf{o}^{\mathbf{f}} - \boldsymbol{\mu}^{\boldsymbol{f}}) \tag{2.12}$$

donde $\boldsymbol{\mu}^{\boldsymbol{nf}}$ y $\boldsymbol{\mu}^{\boldsymbol{f}}$ representan, respectivamente, la media del proceso Gausiano en las dimensiones de las componentes no fiables y fiables; $\boldsymbol{\Sigma}_{\mathbf{nf-f}}$ y $\boldsymbol{\Sigma}_{\mathbf{f-f}}$ son las matrices de covarianzas que expresan, por un lado, las relaciones entre las componentes no fiables y las fiables y, por otro, las relaciones entre las componentes fiables entre sí. Vemos por tanto en la ecuación (2.12), que el valor estimado para las características no fiables es igual a la media de estas características más un término que depende de los puntos fiables que están a su alrededor. Este término mide la distancia de las características fiables a sus medias y la convierte en la distancia a aplicar a las características no fiables (esta conversión se lleva a cabo por medio de las matrices de covarianza).

Al igual que los métodos que modificaban el clasificador, adicionalmente se obtendrían ventajas aprovechando la información disponible sobre las características no fiables.

- Reconstrucción basada en clusters (*Cluster-based reconstruction*): se asume que los vectores de voz limpia se pueden agrupar en clusters, estando cada uno de ellos caracterizado por una distribución Gausiana. Así, la distribución de la voz limpia se corresponde con una mezcla de Gausianas cuyos pesos vienen dados por la probabilidad a priori de que una observación pertenezca a un cluster en particular. Finalmente, la estimación de las componentes no fiables consiste en la suma ponderada de las estimaciones realizadas en cada cluster. Estas estimaciones adoptan una forma similar a la ecuación (2.12) donde las medias y las matrices de covarianza son sustituidas por las representativas de cada cluster.

  Al igual que en los métodos anteriores se obtendrían mejores resultados

empleando la información disponible de las muestras no fiables.

A pesar de que los métodos que permiten reconocer usando parámetros cepstrales son los que mejores prestaciones presentan debemos decir que, cuando todos los sistemas se construyen empleando las log-energías en banda, es el método de marginalización el que obtiene las mejores prestaciones.

En general, estos métodos se muestran tremendamente efectivos cuando se asumen conocidas las regiones del espectrograma que son no fiables. Sin embargo, cuando se emplean clasificadores reales para determinar estas regiones se producen errores en la clasificación que tienen un gran impacto en la robustez del reconocedor. Es por ello que el diseño de clasificadores eficaces [Raj, 2000, Seltzer et al., 2004] se ha vuelto una tarea imprescindible para el éxito de este tipo de técnicas.

Como alternativa a un clasificador explícito, [Barker et al., 2005] proponen buscar de forma conjunta la secuencia de estados y la máscara fiable/no fiable del espectrograma que produce la máxima verosimilitud. Otros autores, como antesala a los métodos basados en la decodificación con incertidumbre (que abordamos en la siguiente sección), en lugar de realizar una clasificación dura en muestras fiables y no fiables, cuantifican el nivel de fiabilidad de las muestras [Barker et al., 2001].

2.6. Métodos basados en decodificación con incertidumbre

Como vimos en la Sección 2.3, las técnicas clásicas de regeneración de parámetros persiguen obtener, a la entrada del reconocedor, coeficientes con la menor cantidad posible de ruido. Sin embargo, estas técnicas no suelen incorporar ningún tipo de información referente a la calidad de esa estimación. Es decir, una vez compensados los parámetros, se asume que éstos ya están libres de cualquier distorsión y se procede con el reconocimiento de habla convencional.

Las técnicas basadas en decodificación con incertidumbre (*uncertainty*

decoding) [Morris et al., 2001, Deng et al., 2002, Arrowood and Clements, 2002, Droppo et al., 2002, Benitez et al., 2004, Stouten et al., 2006] incorporan ese conocimiento en el reconocedor. Estas técnicas asumen que la relación entre los parámetros limpios y los compensados puede modelarse probabilísticamente con el fin de obtener reconocedores más robustos. A continuación veremos cómo esta información es tenida en cuenta estableciendo una comparación entre los reconocedores convencionales y los basados en decodificación con incertidumbre.

Tal y como estudiamos en la Sección 1.1.4, un reconocedor convencional evalúa la probabilidad de que cada modelo HMM haya generado las observaciones y selecciona el de máxima verosimilitud:

$$\lambda = \arg\max_i p(\lambda_i|\mathbf{O}) = \arg\max_i p(\lambda_i)p(\mathbf{O}|\lambda_i) \tag{2.13}$$

siendo λ el modelo ganador; λ_i el modelo candidato i-ésimo y $p(\lambda_i)$ su probabilidad a priori; $\mathbf{O} = \{\mathbf{o_1}, \ldots, \mathbf{o_T}\}$ la secuencia de vectores de parámetros a la entrada del reconocedor, con $\mathbf{o_t}$ el vector observado en el instante t y T el número de vectores observados; por último, $p(\mathbf{O}|\lambda_i)$ representa la probabilidad de que el modelo λ_i haya generado las observaciones $\mathbf{O}$. Finalmente, esta probabilidad se aproxima por la probabilidad de que la secuencia de estados más probable del modelo λ_i haya generado las observaciones. Así, la ecuación (2.13) puede ser reescrita como:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} p(\mathbf{o_t}|x_t^i)\right] p(\lambda_i) \tag{2.14}$$

siendo $\mathbf{X^i} = \{x_0^i, \ldots, x_T^i\}$ la secuencia de estados que produce la máxima verosimilitud en el modelo λ_i; $a_{x_t^i x_{t+1}^i}$ representa la probabilidad de transición del estado x_t^i al x_{t+1}^i para el modelo λ_i; por último, $p(\mathbf{o_t}|x_t^i)$ representa la verosimilitud de que en el estado x_t^i se emita la observación $\mathbf{o_t}$.

Veamos ahora cuál sería el criterio equivalente a la ecuación (2.14) en el caso de aplicar técnicas basadas en decodificación con incertidumbre. Como ya se indicó anteriormente, las técnicas basadas en decodificación con incertidumbre están especialmente diseñadas para el caso en el que a la entrada del reconocedor no tenemos

parámetros limpios sino una estimación de los mismos, $\hat{\mathbf{O}}$. Dicha estimación posee un cierto grado de incertidumbre modelado mediante una distribución de probabilidad,

$$p(\mathbf{O}|\hat{\mathbf{O}}). \tag{2.15}$$

Esta distribución de probabilidad nos indicará la verosimilitud con la que los parámetros limpios $\mathbf{O}$, tras ser contaminados, derivaron en una versión estimada $\hat{\mathbf{O}}$. Debido a la incorporación de esta información se proponen modificaciones en el reconocedor; en particular: en lugar de evaluar la verosimilitud de los modelos en un único punto se evalúa la verosimilitud de todos los valores que pudieron producir aquella estimación. De este modo, el nuevo criterio del reconocedor pasa a formularse del siguiente modo [Morris et al., 2001, Deng et al., 2002, Arrowood and Clements, 2002, Droppo et al., 2002, Benitez et al., 2004, Stouten et al., 2006]:

$$\begin{aligned}\lambda &= \arg\max_i E\left\{p(\lambda_i)p(\mathbf{O}|\lambda_i)\middle|\mathbf{O}\sim p(\mathbf{O}|\hat{\mathbf{O}})\right\} = \\ &= \arg\max_i p(\lambda_i)\int_{\mathbf{O}} p(\mathbf{O}|\lambda_i)p(\mathbf{O}|\hat{\mathbf{O}})d\mathbf{O}\end{aligned} \tag{2.16}$$

Comparando el reconocedor basado en decodificación con incertidumbre, ec. (2.16), con el reconocedor convencional, ec. (2.13), es inmediato observar la diferencia antes reseñada, en lugar de evaluar la distribución de probabilidad para un valor determinado de la observación, $p(\mathbf{O}|\lambda_i)$, ésta es evaluada en un intervalo, definido por $p(\mathbf{O}|\hat{\mathbf{O}})$, como $\int_{\mathbf{O}} p(\mathbf{O}|\lambda_i)p(\mathbf{O}|\hat{\mathbf{O}})d\mathbf{O}$.

Al igual que hicimos con el reconocedor convencional pasamos a continuación a introducir los términos que tienen que ver con la secuencia oculta de estados en la ecuación (2.16). De este modo llegaremos a una expresión donde se ponga de manifiesto la relevancia de cada observación en el cómputo de la verosimilitud final. Para hacerlo debemos primero asumir que la distribución de probabilidad que modela la incertidumbre, $p(\mathbf{O}|\hat{\mathbf{O}})$ en la ec. (2.15), puede escribirse en términos de la distribución de probabilidad de cada vector de observación, $\mathbf{o_t}$, del siguiente modo:

$$p(\mathbf{O}|\hat{\mathbf{O}}) = \prod_{t=1}^{T} p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) \tag{2.17}$$

donde hemos asumido implícitamente que la incertidumbre del vector $\mathbf{o_t}$ es independiente de las observaciones en otros instantes de tiempo.

Teniendo en cuenta la ecuación (2.17) podemos ya introducir la información de la secuencia oculta de estados en la ecuación (2.16):

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i) p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) d\mathbf{o_t} \right] p(\lambda_i) \tag{2.18}$$

Aunque la secuencia de estados oculta que proporciona la máxima verosimilitud para el reconocedor convencional y el modificado no tienen por qué coincidir, comparando la ecuación (2.14) con la ecuación (2.18) vemos que la única diferencia estriba en que el término

$$p(\mathbf{o_t}|x_t^i) \tag{2.19}$$

en el reconocedor convencional pasa a ser

$$\int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i) p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) d\mathbf{o_t} \tag{2.20}$$

en el reconocedor basado en la teoría de decodificación con incertidumbre.

Este tipo de técnicas han sido aplicadas en diferentes sistemas de reconocimiento. Así, se han realizado propuestas definiendo sistemas que actúan sobre varios tipos de parametrizaciones empleando diversos métodos de reducción de ruido. En [Morris et al., 2001, Deng et al., 2002] o [Krisjansson and Frey, 2002] encontramos sistemas construidos usando las log-energías en banda como parametrización. En [Morris et al., 2001] no se emplea ninguna técnica de compensación de ruido y la incertidumbre de las observaciones viene dada por el hecho de que, en presencia de ruidos aditivos, el valor no contaminado asociado a una componente del espectro de potencia nunca supera el valor observado y, por tanto, su incertidumbre viene delimitada a valores comprendidos entre cero y el valor observado. Debemos decir que este trabajo presenta la peculiaridad de que la decodificación con incertidumbre es usada al mismo tiempo que la decodificación convencional ponderando la aportación de una u otra en función del grado de distorsión en las observaciones. Sin embargo,

en [Deng et al., 2002] y [Krisjansson and Frey, 2002] ya se emplean métodos de regeneración de parámetros y la decodificación con incertidumbre es aplicada sobre todas las observaciones. La principal diferencia entre estos métodos estriba en la manera en la que se modela la incertidumbre. En el primer caso, la incertidumbre es modelada a través de una distribución Gausiana mientras que en el segundo caso se inclinan por una estimación de la función de distribución de las observaciones que no se limita a distribuciones Gausianas.

Como es bien sabido, los sistemas basados en las log-energías en banda tienen peores prestaciones que los sistemas basados en parametrizaciones cepstrales. En [Arrowood and Clements, 2002, Droppo et al., 2002, Benitez et al., 2004] o [Stouten et al., 2006] encontramos sistemas basados en parametrizaciones cepstrales que hacen uso de la teoría de decodificación con incertidumbre. De nuevo, la principal diferencia entre unas y otras técnicas radica en los métodos de reducción de ruido empleados.

En cualquier caso, de todos estos trabajos puede deducirse una conclusión clara: no hay ningún método de reducción de ruido ideal e incorporar en el decodificador información acerca de la incertidumbre de los parámetros hace los sistemas más robustos.

2.6.1. Modelado de la incertidumbre

Hasta ahora nada se ha dicho de cómo elegir una distribución de probabilidad adecuada para modelar la incertidumbre sobre las observaciones (parámetros). Para abordar esta elección es interesante tener en cuenta los siguientes principios [Morris et al., 2001]:

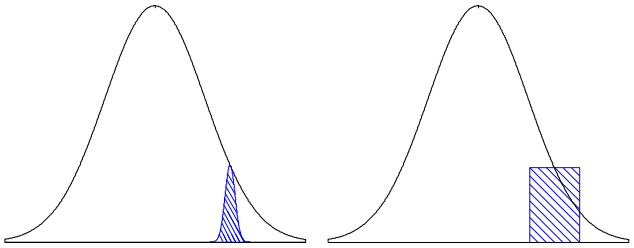
- Hay ocasiones en las que ciertos parámetros de la señal de voz están tan enmascarados por el ruido que no es posible eliminar la distorsión. En estos casos será especialmente importante que el decodificador tenga en cuenta la alta incertidumbre inherente a estos parámetros, ya que su correcta estimación se

torna poco verosímil.

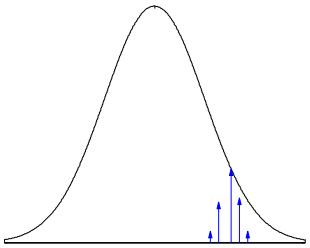
- La distribución de probabilidad que modela la incertidumbre debe ser tan precisa como sea posible, pero sin eliminar ningún posible candidato.

Teniendo en cuenta estas premisas es más sencillo explicar los resultados encontrados por diversos autores para distintas distribuciones de probabilidad. Así, [Morris et al., 2001] emplean distribuciones de probabilidad Uniformes y distribuciones de probabilidad Gausianas encontrando los mejores resultados para las primeras. El rango de hipótesis abarcado por la distribución Uniforme es mayor que el abarcado por la distribución Gausiana, lo que parece ayudar a mejorar los resultados en presencia de ruido. En las Figuras 2.2(a) y 2.2(b) hemos representado de forma ilustrativa los intervalos del espacio de observaciones consideradas por la distribución Gausiana y Uniforme. En dichas figuras, hemos sombreado el rango de valores que serían evaluados al emplear las distintas distribuciones que modelan la incertidumbre (en las figuras, por simplicidad, hemos asumido que los modelos están representados por medio de una sola Gausiana).

Tampoco [Stouten et al., 2006] encuentran sus mejores resultados empleando una distribución Gausiana. Estos autores prueban con distintas distribuciones de probabilidad y encuentran ventajas en emplear una distribución que considera como hipótesis las estimaciones obtenidas por distintos estimadores. En concreto, en este trabajo emplearon como método de compensación de ruido un sistema basado en mezcla de Gausianas y la estimación original consistía simplemente en la suma ponderada de la estimación llevada a cabo por cada Gausiana. Sin embargo, estos autores encuentran mejores resultados cuando, en lugar de pasarle una única estimación al reconocedor, le pasan las llevadas a cabo por cada Gausiana, esto sería equivalente a emplear una distribución discreta para modelar la incertidumbre y así lo representamos en la Figura 2.2(c). En dicha figura, cada delta de Dirac representa la posición de un candidato que será evaluado en la distribución que modela cada estado en los HMM, representada por medio de una sola Gausiana en la figura.



(a) Gausiana. (b) Uniforme.



(c) Conjunto discreto de valores.

Figura 2.2: Ejemplos de distribuciones de probabilidad que modelan la incertidumbre de las observaciones.

Aun así debemos decir que la distribución de probabilidad más comúnmente empleada es la distribución Gausiana [Deng et al., 2002, Arrowood and Clements, 2002, Droppo et al., 2002, Benitez et al., 2004]. Esto en parte se debe a que se obtienen expresiones sencillas para el nuevo criterio empleado en el reconocedor.

2.6.2. Relación entre métodos basados en la decodificación con incertidumbre y métodos basados en las características más fiables

Existe una estrecha relación entre los métodos basados en decodificación con incertidumbre y los que basan el reconocimiento en las características más fiables (*missing-features*), que fueron estudiados en la Sección 2.5. De entre las técnicas que se describieron entonces, marginalización es la que presenta una mayor similitud. El método de marginalización, tal y como expresa la ecuación 2.11, elimina del proceso

de reconocimiento las características no fiables marginalizando las distribuciones de probabilidad conjuntas que representan los modelos y aborda el reconocimiento con las distribuciones marginales resultantes. De este modo, el criterio del reconocedor convencional (ec. (2.14)) pasa a expresarse de la siguiente forma:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \int_{\mathbf{o_t^{nf}}} p(\mathbf{o_t^{nf}}, \mathbf{o_t^f} | x_t^i) d\mathbf{o_t^{nf}} \right] p(\lambda_i) \tag{2.21}$$

siendo $\mathbf{o_t^f}$ el vector con las componentes fiables de $\mathbf{o_t}$ y $\mathbf{o_t^{nf}}$ el vector con las componentes no fiables.

Es interesante recordar aquí que cuando se conocen los valores límites para las características no fiables la integral únicamente se evalúa en los valores posibles de dichas características. De este modo, en lugar de integrar a lo largo de todo el espacio de las observaciones no fiables, se integra únicamente en el rango donde se sabe que pudieron estar las observaciones limpias. Sin duda se está incorporando información acerca de la incertidumbre de las observaciones tal y como lo hacen los métodos basados en la decodificación con incertidumbre. De hecho, existen obvias similitudes entre la ecuación (2.21), que describe el método de marginalización, y la ecuación (2.18), que representa a los métodos basados en la decodificación con incertidumbre. Esta relación todavía se hace más evidente si modelamos la incertidumbre en la ec. (2.18), $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$, mediante una distribución Uniforme. Sin embargo debemos decir que al modelar la incertidumbre por una distribución Uniforme en la ecuación (2.18) aparece un término normalizador que no existe en la ecuación (2.21).

Esta diferencia ya desaparece cuando estudiamos el trabajo de [Morris et al., 2001]. Este trabajo englobado en los métodos que se basan en las características más fiables ya introduce el mismo desarrollo teórico que el empleado en las técnicas basadas en decodificación con incertidumbre y desaparece la diferencia del término normalizador. Estos autores van todavía más allá y, como apuntamos en la sección anterior, no se limitan a distribuciones de probabilidad Uniformes. Aun así, todavía se mantiene la diferencia de clasificar, aunque sea asignando probabilidades en lugar de decisiones duras, los puntos del log-espectro

como fiables o no fiables. De este modo, el criterio empleado por estos autores es el siguiente:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a^i_{x_t^i x_{t+1}^i} \left(w_t p(\mathbf{o_t}|x_t^i) + (1-w_t) \int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i) p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) d\mathbf{o_t} \right) \right] p(\lambda_i) \tag{2.22}$$

siendo w_t el peso que cuantifica la probabilidad de que la observación $\mathbf{o_t}$ sea fiable. Este peso también es función del coeficiente particular en el vector $\mathbf{o_t}$, de modo que habrá componentes dentro de ese vector con distinta probabilidad de estar dominadas por la voz, en la ecuación (2.22) no hemos hecho un énfasis explícito de esta dependencia por motivos de claridad.

Por tanto, podemos interpretar los métodos basados en decodificación con incertidumbre como métodos que se basan en las características más fiables. En este caso, el reconocedor considera como no fiables las características que presentan un alto grado de incertidumbre. Ese alto grado de incertidumbre hace que marginalicemos sus funciones de densidad de probabilidad eliminando su efecto en el reconocedor.

Por otro lado, la aplicación de estas técnicas basadas en las características más fiables todavía presenta el problema de emplear las log-energías en bandas para diseñar los reconocedores. En el trabajo presentado en esta tesis este inconveniente es obviado gracias al empleo de la parametrización FF (*Frequency Filtered*, [Nadeu et al., 1995, Nadeu et al., 2001, Paliwal, 1999]). Estos parámetros, además de conseguir prestaciones equivalentes a los parámetros cepstrales, tienen la ventaja de permanecer en el dominio de la frecuencia lo que nos permite analizar y extraer conclusiones más intuitivas sobre las bondades de los métodos propuestos.

En el Capítulo 4, presentaremos el desarrollo de un sistema robusto basado en la teoría de decodificación con incertidumbre para la parametrización FF.

Capítulo 3

Reconocimiento robusto en sistemas RAH por medio de la combinación de bounded-distance HMM y sustracción espectral

3.1. Introducción

En el Capítulo 2 revisamos algunas de las técnicas surgidas para aumentar la robustez de los sistemas de reconocimiento automático de habla (RAH). Estas técnicas trataban de compensar principalmente el efecto de las distorsiones convolutivas y de los ruidos aditivos. En este capítulo, nos centraremos en estos últimos, los ruidos aditivos.

Así, proponemos emplear una técnica que denominamos *bounded-distance HMM* para mitigar el efecto de estos ruidos aditivos. Esta técnica fue usada por [Matsui and Furui, 1992] en un sistema de reconocimiento de locutor y nosotros la aplicaremos a un sistema de reconocimiento automático de habla. *Bounded-distance HMM* es capaz de reducir el impacto que tienen las características que son *outliers*

para los modelos acústicos en el sistema de reconocimiento. Es decir, reduce el impacto de las características claramente corruptas y que, por tanto, no están bien representadas por las distribuciones estadísticas aprendidas en la fase de entrenamiento.

La aplicación de esta técnica lleva implícita la detección de los *outliers*. De este modo, esta técnica nos recuerda a las técnicas estudiadas que basan el reconocimiento en las características fiables descartando el resto (*missing-features*, ver Sección 2.5). Las similitudes entre este tipo de técnicas y *bounded-distance HMM* son patentes y, a lo largo de este capítulo, realizamos un breve análisis que nos permite englobar esta técnica en aquella categoría.

Por otro lado, encontramos otra técnica similar conocida como *acoustic backing-off* [de Veth et al., 2001a] que se aplicó a sistemas de reconocimiento automático de habla. *Acoustic backing-off* modifica la distribución de probabilidad de los modelos con el fin de adaptarlos a las características de entrada que no están bien representadas por los datos de entrenamiento. De este modo, proponen sumar una distribución uniforme a la distribución generada a partir de los datos de entrenamiento. Así, se consigue que estas características para las cuales no teníamos datos de entrenamiento no tengan un impacto severo sobre el proceso de reconocimiento. El problema principal de esta técnica es que consigue prestaciones pobres para ruidos de banda-ancha [de Veth et al., 2001b].

En este capítulo proponemos combinar *bounded-distance HMM* y sustracción espectral para superar los problemas que presenta *acoustic backing-off* ante este tipo de ruidos. Como veremos a lo largo del capítulo, tanto sustracción espectral como *bounded-distance HMM* se complementan mutuamente. Por un lado, el efecto de las distorsiones no lineales introducidas por sustracción espectral resulta adecuadamente tratado por el método *bounded-distance HMM* y, por otro lado, la compensación de características llevada a cabo por sustracción espectral resulta beneficiosa para *bounded-distance HMM*.

En este capítulo, mostramos, tanto teórica como experimentalmente, las ven-

tajas de la combinación propuesta. En concreto, esta combinación ha sido satisfactoriamente evaluada para diversas tareas de reconocimiento donde estudiamos el efecto de un gran número de ruidos para distintas relaciones señales a ruido [Vicente-Peña et al., 2007].

El resto de las secciones de este capítulo se organizan del siguiente modo. La Sección 3.2 introduce el método *bounded-distance HMM*. En la Sección 3.3 detallamos nuestra propuesta basada en la combinación de métodos donde se incluyen algunos detalles específicos de nuestra implementación. A continuación, mostramos los experimentos y resultados en la Sección 3.4. Por último, en la Sección 3.5 presentamos las principales conclusiones obtenidas a partir de nuestros resultados.

3.2. Bounded-distance HMM

3.2.1. Motivación y trabajos previos

En el Capítulo 1 estudiamos el funcionamiento de un reconocedor automático de habla basado en HMMs. Vimos que, dada una secuencia de observaciones, para determinar el mensaje lingüístico contenido en la señal de voz, se evalúa la log-verosimilitud de cada una de las unidades acústicas consideradas y se selecciona la que proporciona una verosimilitud máxima. Así, esta decisión puede expresarse matemáticamente como:

$$\lambda = \arg\max_i \left(\log(a_{x_o^i x_1^i}) + \sum_{t=1}^{T} \left[\log(p(\mathbf{o_t}|x_t^i)) + \log(a_{x_t^i x_{t+1}^i}) \right] + \log(p(\lambda_i)) \right) \quad (3.1)$$

donde:

- λ_i es el i-ésimo modelo acústico y λ es el modelo ganador.
- $a_{x_t^i x_{t+1}^i}$ es la probabilidad de transición entre los estados x_t^i y x_{t+1}^i para el modelo λ_i.
- $\mathbf{o}_t$ es el vector de características en el instante t.

- $p(\mathbf{o_t}|x_t^i)$ es la probabilidad de emisión del vector de características $\mathbf{o_t}$ en el estado x_t^i del modelo λ_i
- T es el número de vectores de características de entrada.
- $p(\lambda_i)$ es la probabilidad "a priori" del modelo λ_i.

Para introducir el método *bounded-distance HMM* (BD-HMM) nos fijamos ahora en la log-probabilidad de emisión de las observaciones, $\log(p(\mathbf{o_t}|x_t^i))$. Asumiendo que esta log-probabilidad está modelada por una sola Gausiana, este término se puede reescribir como:

$$\log(p(\mathbf{o_t}|x_t^i)) = -\frac{1}{2}\log\left((2\pi)^N|\boldsymbol{\Sigma}_{\boldsymbol{x_t^i}}|\right) - \frac{1}{2}(\mathbf{o_t} - \boldsymbol{\mu}_{\boldsymbol{x_t^i}})^T(\boldsymbol{\Sigma}_{\boldsymbol{x_t^i}})^{-1}(\mathbf{o_t} - \boldsymbol{\mu}_{\boldsymbol{x_t^i}}) \tag{3.2}$$

donde N es la dimensión del vector de características de entrada y $\boldsymbol{\mu}_{\boldsymbol{x_t^i}}$ y $\boldsymbol{\Sigma}_{\boldsymbol{x_t^i}}$ representan, respectivamente, la media y la matriz de covarianzas de la Gausiana que representa el estado x_t^i del modelo λ_i.

Prescindiendo en la ecuación (3.2), por simplicidad, de la notación que alude al tiempo, al estado actual y al modelo y, por otro lado, asumiendo matrices de covarianza diagonales obtenemos:

$$\begin{aligned}\log(p(\mathbf{o_t}|x_t^i)) &= -\frac{1}{2}\log\left((2\pi)^N\prod_{k=1}^{N}\sigma_k^2\right) - \frac{1}{2}\sum_{k=1}^{N}\frac{(o_k-\mu_k)^2}{\sigma_k^2} = \\ &= -\frac{1}{2}\left\{\sum_{k=1}^{N}\log(2\pi\sigma_k^2) + \sum_{k=1}^{N}\frac{(o_k-\mu_k)^2}{\sigma_k^2}\right\}\end{aligned} \tag{3.3}$$

donde σ_k^2 representa la k-ésima componente de la diagonal de la matriz de covarianzas y μ_k la k-ésima componente del vector de medias.

A partir de la ecuación (3.3) queda claro que esta log-probabilidad depende de la distancia euclídea normalizada entre la observación actual y la media del modelo. En concreto, el término

$$-\frac{1}{2}\sum_{k=1}^{N}\frac{(o_k-\mu_k)^2}{\sigma_k^2} \tag{3.4}$$

dominará la expresión de la ecuación (3.3) si una de las componentes del vector de características, o_k, está altamente contaminada y, por tanto, se sitúa a gran distancia de su media, μ_k. De este modo, esta componente altamente contaminada contribuye al descarte del modelo representado por dicho vector de medias. Por ello resulta conveniente limitar esta distancia y, así, la influencia de estos *outliers* en la decisión final. De esta manera, la distancia euclídea normalizada de la ecuación (3.4) se sustituye por otra que adopta la siguiente expresión:

$$-\frac{1}{2}\sum_{k=1}^{N}\frac{BD(o_k-\mu_k)}{\sigma_k^2} \tag{3.5}$$

donde

$$BD(o_k-\mu_k)=\begin{cases}(o_k-\mu_k)^2, & \text{si } |o_k-\mu_k|<\alpha\sigma_k\\ (\alpha\sigma_k)^2, & \text{resto}\end{cases} \tag{3.6}$$

donde α es un parámetro que controla el valor del límite. En la Figura 3.1 hemos representado la distancia euclídea original y la distancia acotada que aquí usaremos. En caso de tener, en lugar de una sola Gausiana, una mezcla de Gausianas para representar nuestro modelo, aplicaríamos la ecuación (3.6) a cada componente en la mezcla. Por otro lado, debemos destacar que, al emplear esta limitación en la distancia euclídea, los estados ya no están representados por una función de densidad de probabilidad ya que su integral en el espacio de las características de entrada produce un valor mayor que uno.

La idea subyacente a BD-HMM no es nueva y ya fue aplicada a sistemas de reconocimiento de locutor [Matsui and Furui, 1991, Matsui and Furui, 1992]. [Matsui and Furui, 1991] emplean una nueva distancia denominada DIM (*Distortion-Intersection Measure*) en un sistema de reconocimiento de locutor basado en cuantificación vectorial. Posteriormente, esta distancia se adapta para ser usada en sistemas de reconocimiento de locutor basados en HMM [Matsui and Furui, 1992]. En este último trabajo, la cola de las Gausianas que representan los modelos de cada locutor se aplana para así limitar la verosimilitud para las observaciones que se sitúan muy lejos de la media. En concreto, se limita la verosimilitud de las observaciones que

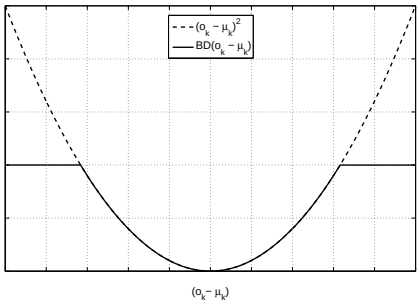


Figura 3.1: Distancia euclídea acotada (trazo continuo) frente a la distancia euclídea (trazo discontinuo).

caen a una distancia de la media superior a 3 desviaciones típicas. Este límite es equivalente a asignar $\alpha = 3$ en la ecuación (3.6).

En el ámbito del reconocimiento automático de habla encontramos una propuesta conocida como *acoustic-backing off* [de Veth et al., 1998, de Veth et al., 2001a, de Veth et al., 2001b] con importantes similitudes a BD-HMM. Los autores que proponen esta técnica defienden que la distribución Gausiana que habitualmente se emplea en los sistemas HMM no es suficiente para representar las características de entrada que no han sido vistas en la etapa de entrenamiento. Así, consideran que la distribución Gausiana obtenida a través de la etapa de entrenamiento es únicamente adecuada para las características que están bien representadas por el conjunto de entrenamiento mientras que no representa adecuadamente al resto de características. Por tanto, proponen emplear una distribución compuesta por la suma ponderada de dos distribuciones: la distribución Gausiana estimada a través de la etapa de entrenamiento y una distribución uniforme (ya que no disponemos de conocimiento "a priori") que trata de representar a las características que no fueron observadas en dicha etapa de entrenamiento. El resultado es una distribución similar a una Gausiana que en lugar de tender a cero según nos alejamos de la media se satura a un determinado valor (dado por el nivel de la distribución uniforme).

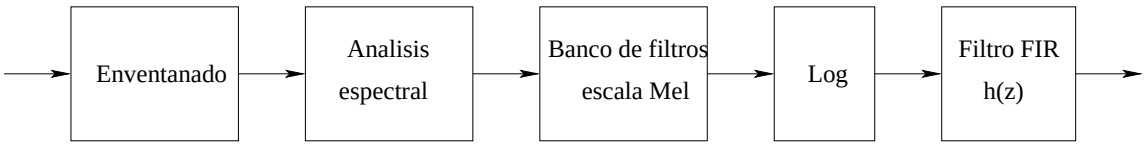


Figura 3.2: Diagrama de bloques de la parametrización FF

Siempre que la distribución que modela las características no observadas en la etapa de entrenamiento se aproxime mediante una distribución uniforme, *acoustic-backing off* es similar a BD-HMM. El papel desempeñado por el parámetro α en la ecuación (3.6) es ahora desempeñado por el nivel de la distribución uniforme. No obstante, [de Veth et al., 2001a] también modifican la función de densidad de probabilidad para las muestras que no son consideradas *outliers*.

La eficacia de *acoustic-backing off* depende de la parametrización empleada [de Veth et al., 2001b]. En concreto, este método es mucho más efectivo para parametrizaciones, como Frequency Filtered (FF), que no propagan una distorsión localizada a todos los coeficientes de la parametrización.

En la Figura 3.2 presentamos el diagrama de bloques utilizado para obtener la parametrización FF [Nadeu et al., 1995, Nadeu et al., 2001]. Como podemos apreciar en la figura, la principal diferencia respecto a la parametrización MFCC (Figura 1.3, Sección 1.1.1) consiste en la sustitución de la DCT por un filtro FIR y que la etapa de liftering ha desaparecido.

El filtro FIR se diseña con el objetivo de cumplir dos cometidos: por un lado los parámetros a su salida deben estar tan decorrelacionados como sea posible y, por otro, debe favorecer los coeficientes más discriminativos. Atendiendo a este último objetivo, se diseña un filtro que, en el dominio cepstral, es equivalente a un liftering que realza los coeficientes cepstrales más relevantes. Finalmente, el filtro

$$h(z) = z - z^{-1} \tag{3.7}$$

cumple ambos propósitos [Paliwal, 1999, Nadeu et al., 2001] y será el que empleemos a lo largo de esta tesis. Por tanto, cada coeficiente FF se calcula como la diferencia

entre dos log-energías en banda lo que hace que permanezcamos en dicho dominio. Si estudiamos el efecto de este filtro en el dominio cepstrum vemos que es equivalente a atenuar los coeficiente cepstrales menores y mayores.

Los parámetros MFCC estiman la envolvente espectral desechando los parámetros cepstrales que provienen de la excitación de la señal de voz y, de este modo, únicamente retienen aquellos coeficientes que representan la envolvente espectral. Esta etapa de liftering que selecciona estos coeficientes es necesaria debido a que, si usamos un número elevado de filtros en el banco de filtros, la estructura fina del espectro todavía esta presente en el dominio de las log-energías en banda. Sin embargo, el número de filtros en dicho banco suele ser inferior para la parametrización FF y, por tanto, en las log-energías en banda ya no se aprecia el efecto de la excitación.

Debido al filtro paso alto empleado en la obtención de los parámetros FF, una distorsión que estuviese claramente localizada en el dominio del espectro y que únicamente afectase a una log-energía en banda se propagaría únicamente a dos coeficientes de la parametrización final. Sin embargo, con la parametrización MFCC, al realizar la DCT a las log-energías en banda propagamos este tipo de distorsiones localizadas a todos los coeficientes. Teniendo en cuenta esta última conclusión, en todos los experimentos realizados en este capítulo se emplea la parametrización FF.

[de Veth et al., 2001b] evalúan las prestaciones de *acoustic-backing off* para tres ruidos aditivos: ruido de coches, de voces y de fábrica extraídos de la base de datos NOISEX-92 [Varga et al., 1992]. Encuentran que su método es efectivo para el ruido de coches, un ruido de banda estrecha y altamente localizado, mientras que no es efectivo para el resto. Afirman que esto se debe a que el ruido de voces y de fábrica son ruidos de banda ancha mientras que el ruido de coches está claramente coloreado. Como veremos más adelante, la combinación de BD-HMM y sustracción espectral supera esta limitación consiguiendo resultados satisfactorios para estos tres tipos de ruido.

3.2.2. BD-HMM y métodos basados en las características más fiables

[de Veth et al., 2001a] interpretan *acoustic-backing off* como una técnica que basa el reconocimiento en las características más fiables (*Missing Feature*, MF) y, del mismo modo, BD-HMM puede también interpretarse dentro de esa filosofía. Los métodos basados en MF [Cooke et al., 2001, Raj et al., 2004] fueron explicados en la Sección 2.5, allí vimos que estos métodos detectan las características no fiables e impiden que tomen parte en el proceso de reconocimiento. De esta manera, el reconocimiento sólo se basa en las características de los vectores de parametrización que se consideran fiables. Aunque estos métodos obtienen prestaciones altas cuando la clasificación de las características entre fiable/no fiable es precisa, sus prestaciones decaen rápidamente cuando se introducen errores en dicha clasificación.

Al igual que MF, BD-HMM aborda el reconocimiento minimizando la influencia que tienen las características no fiables (*outliers*) en el reconocimiento. Sin embargo, la clasificación entre *outliers* (no fiable) y no *outliers* (fiables) se realiza en cada modelo acústico y, por tanto, no es necesario el diseño explícito de tales clasificadores. Una característica es considerada outlier si su distancia a la media del modelo que actualmente evaluamos es mayor que un determinado umbral, $\alpha\sigma_k$ en la ecuación (3.6).

No obstante, debemos destacar una diferencia entre MF y BD-HMM. BD-HMM únicamente aborda la detección de *outliers*, es decir, características que claramente son no fiables, mientras que MF va más allá y aborda la tarea más compleja de decidir si las características de entrada son fiables o no fiables.

3.3. Combinación de bounded-distance HMM y sustracción espectral

3.3.1. Sustracción Espectral

Sustracción espectral (*spectral subtraction*, SS) fue estudiada en la Sección 2.3.1 y aquí únicamente detallamos las particularidades de nuestra implementación.

Como estimador de la potencia del espectro de ruido hemos empleado el método que propone [Martin, 2001]. Básicamente, este método busca los mínimos en la densidad espectral de potencia de la señal de voz contaminada y los relaciona con las características del ruido. En concreto, asumiendo que los instantes de tiempo donde se producen esos mínimos se corresponden con silencios donde únicamente está presente la señal de ruido, esos mínimos en el espectro de potencia están relacionados mediante un factor de proporcionalidad con la media de la densidad espectral de potencia del ruido.

Nuestra implementación de SS está basada en diferentes ideas que aparecen en la literatura. En concreto, nos basamos en las siguientes ideas:

- Antes de realizar la sustracción de la densidad espectral de potencia del ruido, [Boll, 1979] propone realizar un suavizado de la densidad espectral de potencia de la señal contaminada. Es importante que este suavizado no modifique la inherente variabilidad de la señal de voz por lo que no conviene que se expanda entre demasiadas muestras. De este modo, realizamos una media entre la densidad espectral de potencia de las tramas anterior, actual y posterior. Si asumimos que la señal de ruido es estacionaria a lo largo de estas tres tramas, estamos reduciendo la varianza del ruido por un factor 3.

- Es más conveniente aplicar SS antes del bando de filtros Mel [Nolazco Flores and Young, 1994]. Si, por el contrario, se aplica a la salida del bando de filtros, las componentes en frecuencia que no serían correctamente estimadas mediante SS (ya que su estimación daría valores

negativos de la potencia del espectro) quedarían enmascaradas por el resto de componentes en frecuencia de su misma banda.

- La sustracción llevada a cabo en SS podría producir valores del espectro de potencia negativos por lo que debemos decidir qué hacer para evitar estos valores. En nuestra implementación imponemos un valor mínimo a la potencia del espectro que es proporcional a la densidad espectral de la potencia del ruido. Esta solución se usa habitualmente y un ejemplo reciente lo encontramos en [Pujol et al., 2004].

A modo de resumen, nuestra implementación de SS queda descrita por la siguiente ecuación[1]:

$$\widehat{P_X}(\omega, l) = \max\left\{\widehat{P_S}(\omega, l) - \gamma\widehat{P_N}(\omega, l), \beta\widehat{P_N}(\omega, l)\right\} \tag{3.8}$$

donde $\widehat{P_X}$, $\widehat{P_S}$ y $\widehat{P_N}$ son, respectivamente, las estimaciones de la potencia del espectro para la voz limpia, la voz contaminada y el ruido; ω es el índice frecuencial y l el temporal; γ y β son las constantes que introdujimos en la Sección 2.3.1 denominadas "factor de sobreestimación" y "nivel de suelo". Tal y como indicamos anteriormente, $\widehat{P_S}$ se estima realizando la media de las potencias del espectro de las tramas anterior, actual y posterior. Finalmente, incidimos en que $\widehat{P_N}$ se estima empleando el método que propone [Martin, 2001].

Como ya apuntamos en la Sección 2.3.1, a pesar de que SS es capaz de aumentar la SNR de la señal de voz, sus no linealidades producen distorsiones que pueden degradar las prestaciones de los reconocedores automáticos de habla [Gong, 1995]. En la siguiente sección vemos como BD-HMM complementa perfectamente a SS de modo que se reducen los efectos negativos de estas distorsiones.

[1]Esta ecuación es similar a la ecuación (2.6) pero la hemos reescrito aquí por conveniencia

3.3.2. Combinación de bounded-distance HMM y sustracción espectral

BD-HMM es efectivo para reducir el impacto de los *outliers* cuando reconocemos voz contaminada. Como ya explicamos en la Sección 3.2, BD-HMM impone un límite a la distancia euclídea que aparece en el exponente de las Gausianas de modo que los *outliers* no dominen esa distancia y, así, se reduce su impacto en los reconocedores. Sin embargo, BD-HMM únicamente actúa sobre los *outliers* dejando invariable el resto de efectos que produce el ruido.

Por otro lado, SS trata de estimar los parámetros limpios a partir de los contaminados y, a diferencia de BD-HMM, actúa sobre todo el espacio de características de entrada. Sin embargo, SS no fue originalmente diseñado como una etapa previa al reconocimiento sino como una técnica de reducción de ruido. En este contexto, se sabe que esta reducción de ruido se consigue a costa de introducir distorsiones en lo parámetros. Como experimentalmente constatamos en la Sección 3.4 estas distorsiones no lineales aumentan el número de *outliers*. Teniendo en cuenta estos aspectos relativos a ambos métodos, proponemos el empleo conjunto de SS y BD-HMM. De este modo, nos beneficiamos del poder reductor de ruido de SS mientras que, por medio de BD-HMM, reducimos el efecto de los *outliers* que genera SS. Además, BD-HMM mejora sus prestaciones gracias a SS ya que este último método compensa todos los parámetros y no sólo los *outliers*.

3.3.3. Detalles de nuestra implementación

Antes de presentar los resultados que apoyan la combinación de ambos métodos, es necesario clarificar algunos aspectos relativos a su implementación.

El único parámetro libre del método BD-HMM es el parámetro α en la ecuación (3.6). Este parámetro determina qué características de entrada son consideradas *outliers*. En nuestra implementación hemos asignado el valor $\alpha = 3$ que coincide con el que emplean [Matsui and Furui, 1992]. El límite impuesto para este valor de α

no degrada las prestaciones de los sistemas en ausencia de *outliers* ya que, al tratar con distribuciones Gausianas, el 99.7 % de las muestras caen a una distancia de la media comprendida en $\pm 3\sigma$ (los resultados que demuestran experimentalmente esta afirmación se presentan en la Tabla 3.2, Sección 3.4).

En cuanto a los parámetros libres de SS, debemos configurar los valores de γ y β (ec. (3.8)). Estos se determinaron por medio de un pequeño barrido que ha considerado los siguientes valores: $\gamma = \{0{,}8;\ 1{,}0\}$ y $\beta = \{0{,}1,\ 0{,}2,\ 0{,}3\}$ para cada ruido y SNR de la tarea RM1 (esta tarea se describirá en detalle en la Sección 3.4). A pesar de que no hubo ningún par de parámetros que se pudiese considerar óptimo para todas las condiciones, el par $\{\gamma = 0{,}8;\ \beta = 0{,}2\}$ parece ser el más adecuado cuando SS se ejecuta de forma aislada. Sin embargo, cuando SS se combina con BD-HMM, el mejor par resultó ser $\{\gamma = 1{,}0;\ \beta = 0{,}1\}$ (en la Sección 3.4.4 se muestran los resultados obtenidos con estos dos pares de parámetros). Por último, estos valores de los parámetros para la tarea RM1 se extrapolan al resto de tareas que son descritas en la Sección 3.4.

3.4. Experimentos y resultados

Para evaluar nuestra propuesta se realizan dos tipos de experimentos. Los primeros sirven para motivar nuestra metodología, en ellos se mide la influencia de los *outliers* en el proceso de reconocimiento. En los segundos se cuantifican las mejoras de nuestra propuesta mediante la tasa de error por palabras (*Word Error Rate*, WER).

Los experimentos se han realizado sobre 4 tareas de reconocimiento automático de habla. En la subsección siguiente presentamos la configuración que las 4 tareas comparten mientras que, en la siguiente, se describe cada tarea con mayor nivel de detalle haciendo hincapié en sus particularidades. Finalmente, presentamos los resultados para los dos tipos de experimentos considerados.

3.4.1. Configuración común a las tareas de reconocimiento

Como indicamos anteriormente, nuestros experimentos se realizan sobre 4 tareas diferentes. Cada una de ellas se diseña apoyándonos en bases de datos conocidas: RM1 [NIST, 1992], Wall-Street Journal [Paul and Baker, 1992], Aurora-4 [Hirsch, 2002a] y Spanish SDC-Aurora [Macho, 2000]. Para cada tarea se construye un sistema de reconocimiento automático de habla base empleando la utilidad HTK [Young et al., 2002].

La misma parametrización se emplea para todos los experimentos. En concreto, el vector de características consiste en 12 parámetros FF más la log-energía que se calculan cada 10 ms empleando ventanas Hamming de 25 ms. La media temporal de los parámetros FF a lo largo de cada locución se elimina y se normaliza la log-energía. Por último, al vector de características estáticas se le añaden los coeficientes de regresión de primer y segundo orden de modo que la dimensión final de nuestro vector de características es igual a 39. Tal y como propone [Shozakai et al., 1997], el conjunto de entrenamiento en cada una de las tareas también es procesado empleando SS.

3.4.2. Detalles específicos de cada tarea. Descripción de las bases de datos

Resource Management RM1 [NIST, 1992]

La base de datos *Resource Management* RM1 tiene un vocabulario de 991 palabras. Nosotros hemos empleado los datos que permiten realizar experimentos independientes de locutor, en ellos se establecen dos grupos: el grupo de entrenamiento que tiene 3.990 frases provenientes de 109 locutores y, el grupo de test que contiene 1.200 frases de 40 locutores y que se corresponden con la recopilación de los cuatro primeros conjuntos oficiales de test (*February 1989, October 1989, February 1991, September 1992*). Para la trascripción de los datos se emplea el diccionario SRI Resource Management (proporcionado en la misma distribución por el NIST) que ha

sido modificado para adaptarlo al conjunto de fonemas propuestos por CMU tal y como se propone en la tarea RM definida en HTK. Para nuestro experimentos empleamos una versión submuestreada de la base de datos a 8 KHz, la base de datos original fue capturada a 16 KHz.

Se utilizan modelos dependientes de contexto (*cross word* trifonemas) con 3 estados y 3 mezclas de Gausianas por estado. Se emplean dos modelos para representar el silencio, uno largo y otro corto. Finalmente, se emplea una bigramática como modelo de lenguaje.

Se generan artificialmente versiones contaminadas del grupo de test de esta base de datos. Estas se generan para 5 tipos de ruido con 4 SNRs diferentes. En concreto, se emplea una versión submuestreada a 8 KHz de los ruidos blanco, rosa, de coches, de voces y de fábrica de la base de datos NOISEX-92 [Varga et al., 1992]. Las SNRs van desde 0 dB hasta 15 dB en pasos de 5 dB. Estas versiones contaminadas sólo se usan para la fase de test y nunca para la de entrenamiento.

Wall Street Jounal (WSJ0) [Paul and Baker, 1992]

La base de datos *Wall Street Journal* (WSJ0) se empleó para los siguientes experimentos. El grupo de entrenamiento estándar *SI-84*, que contiene 7138 frases, se empleó para la construcción de los modelos. Para la fase de test se empleó el conjunto de test *Nov'92 CSR Speaker Independent 5K Read NVP (Non Verbalization Punctuation)* con 330 frases. Al igual que con la base de datos anterior, se empleó una versión submuestreada a 8 KHz.

El diccionario distribuido por CMU [CMU, 1998] se utilizó para obtener la transcripción de las locuciones, para ello se eliminó el acento de las vocales por lo que se emplearon 39 fonemas en nuestras transcripciones. De nuevo, se usaron modelos dependientes de contexto (*cross-word* trifonemas) y dos modelos para representar el silencio. Se utilizaron tres estados por modelo y 8 Gausianas para representar la distribución de cada estado. La única excepción ocurre para los modelos de silencio donde se usan 16 Gausianas. Además, se emplea el modelo de lenguaje de 5000

palabras que se distribuye con la base de datos.

Por último, se crean versiones contaminadas de esta base de datos siguiendo el mismo procedimiento que se empleó con la base de datos RM1.

Aurora-4 [Hirsch, 2002a]

Esta base de datos se basa en la WSJ0 que explicamos en la subsección anterior. Así, el conjunto de entrenamiento coincide con el de la subsección anterior mientras que, el conjunto de test, basándose también en el mismo (*Nov'92*), incluye versiones contaminadas que se distribuyen junto con la base de datos. Estas versiones contaminadas se crearon añadiendo un tipo de ruido bajo una SNR aleatoria entre 5 y 15 dB con pasos de 1 dB. Se emplean seis tipos de ruido diferentes: coches, voces, restaurante, calle, aeropuerto y estación de tren. En nuestros experimentos usamos todos los ruidos pero nos limitamos a las versiones submuestreadas a 8 KHz y al micrófono de solapa.

El sistema de reconocimiento se diseñó exactamente igual que el de la WSJ0.

Spanish SDC-Aurora [Macho, 2000]

A diferencia de las bases de datos anteriores, en la base de datos *Spanish SDC-Aurora* el ruido no se añade artificialmente sino que la señal de voz se captura directamente en un entorno adverso.

Esta base de datos está compuesta por 4.914 frases capturadas tanto con un micrófono de solapa como con un micrófono lejano. Las grabaciones se realizan en tres entornos: *quiet* (dentro del coche con el motor apagado), *low* (conduciendo a baja velocidad en una carretera de una ciudad) y *high* (conduciendo a alta velocidad en una carretera buena). Las grabaciones se capturan a 8 KHz. Se realizan los experimentos estándares propuestos en la base de datos: *Well-Matched* (WM, condiciones de entrenamiento y test similares), *Medium-Mismatch* (MM, ligeras diferencias entre las condiciones de entrenamiento y de test) y *High-Mismatch* (HM, diferencias más notables entre entrenamiento y test). Las condiciones exactas de estos experimentos

Tabla 3.1: Resumen de las principales diferencias entre las 4 tareas de reconocimiento

	RM1	**WSJ0**	**Aurora-4**	**Spanish SDC-Aurora**
Número de palabras	991	5000	5000	10
Tipos de distorsion	Contaminadas para nuestros experimentos: 5 ruidos, 4 SNRs.	Contaminadas para nuestros experimentos: 5 ruidos, 4 SNRs.	6 ruidos bajo una SNR aleatoria	1 ruido real, 3 condiciones de ruido

están detalladas en [Macho, 2000].

El sistema de reconocimiento automático de habla se construye empleando las rutinas distribuidas con la base de datos. Para cada dígito se construyen modelos de 18 estados empleando una mezcla de 3 Gausianas por estado. Al igual que en las bases de datos anteriores, se distinguen dos modelos de silencio que modelan pausas cortas y largas entre los dígitos. El modelo de silencio largo tiene 3 estados modelados por 6 Gausianas mientras que, el silencio corto, únicamente tiene un estado que está asociado al estado central del silencio largo.

En la Tabla 3.1 mostramos un resumen con las principales diferencias entre las 4 tareas que acabamos de explicar.

3.4.3. Influencia de los outliers en los reconocedores

En estos primeros experimentos medimos el porcentaje de *outliers* que están presentes en el proceso de reconocimiento cuando la señal de voz se contamina con ruido aditivo. Para estos experimentos hemos empleado la base de datos RM1 bajo

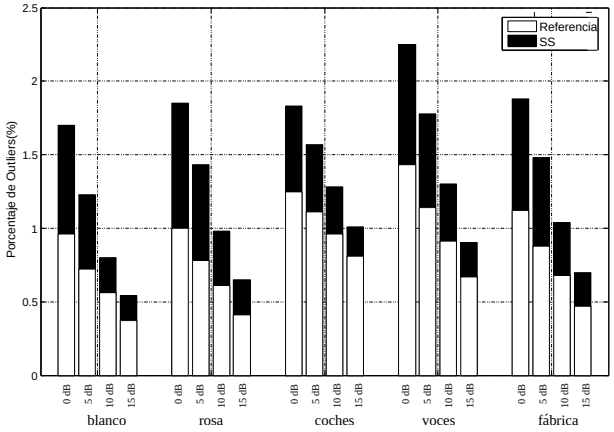


Figura 3.3: Porcentaje medio de *outliers* para el grupo de test de la base de datos RM1 y para varios ruidos y SNRs. Las barras etiquetadas como *Referencia* indican el porcentaje medio calculado sobre los parámetros extraídos a partir de la voz contaminada original. Las barras etiquetadas como *SS* muestran el porcentaje medio cuando se aplica sustracción espectral.

dos situaciones: aplicando y sin aplicar SS. La decisión de que una característica de entrada sea o no un outlier depende del modelo acústico que estemos evaluando o, dicho de otro modo, depende de la secuencia de estados que se atraviesan en el reconocedor. Por ello, en estos experimentos el porcentaje de *outliers* lo calculamos atravesando la secuencia de estados que producen un alineamiento forzado usando las transcripciones correctas y los parámetros extraídos a partir de la voz limpia.

La Figura 3.3 muestra los resultados obtenidos. Las barras blancas indican el porcentaje empleando los parámetros extraídos directamente de la voz original mientras que, las barras negras, representan el porcentaje obtenido cuando se aplica SS a la señal de entrada (en estos experimentos se usó el par de parámetros $\{\gamma = 1{,}0; \beta = 0{,}1\}$). Como se observa en la figura, el número de *outliers* es sistemáticamente mayor cuando se aplica SS. Además y como era de esperar, este porcentaje aumenta cuando disminuye la SNR. Es interesante destacar que estos porcentajes son bajos sin embargo, tal y como mostramos en el siguiente experimento, su influencia en el reconocedor es muy significativa.

Cuando a la entrada de un reconocedor introducimos una señal cuyo mensaje es desconocido, se evalúa la log-probabilidad a lo largo de todos los posibles caminos y se escoge el máximo. Ahora nos concentramos en la parte de esta log-probabilidad que modela lo bien que las observaciones se ajustan a las distribuciones que representan los estados dentro de los HMMs. De este modo, nos referimos a la log-probabilidad acumulada de la ecuación (3.1) como:

$$\log(p)_{accum} = \sum_{t=1}^{T} \log(p(\mathbf{o_t}|x_t^i)) \tag{3.9}$$

Para evaluar la influencia de los *outliers* en la decisión del reconocedor, hemos medido la contribución de estos *outliers* a esta log-probabilidad acumulada. De nuevo, esta log-probabilidad acumulada depende de la secuencia de estados que atravesemos en el reconocedor, por lo que, al igual que antes, este experimento lo realizamos a través de un alineamiento forzado. Así, hemos calculado el término $\log(p)_{accum}$ en dos situaciones: cuando aplicamos BD-HMM y cuando no lo hacemos. Denotamos con $\log(p)_{accum}$ la log-probabilidad acumulada cuando no aplicamos BD-HMM y $\log(p)_{accum}^{BD-HMM}$ esta log-probabilidad acumulada cuando se emplea BD-HMM. Finalmente, con el objetivo de cuantificar la contribución de los *outliers* a la log-probabilidad se calcula el siguiente porcentaje:

$$D(\,\%) = 100\frac{\log(p)_{accum}^{BD-HMM} - \log(p)_{accum}}{\sum_{t=1}^{T} |\log(p(\mathbf{o_t}|x_t^i))|} \tag{3.10}$$

Para hacer este experimento más claro e intuitivo, hemos modificado ligeramente el método BD-HMM para enfatizar el concepto de outlier. En concreto, sólo limitamos la distancia euclídea, tal y como indica la ecuación (3.5), cuando la característica está fuera del alcance de todas las Gausianas que forman la mezcla de Gausianas. En el método BD-HMM original, la distancia euclídea se limita para todas las Gausianas de forma independiente al resto de componentes en la mezcla. De este modo, todas las Gausianas clasifican cada muestra como *outlier*/no *outlier*. Teniendo en cuenta que una característica es un *outlier* para *todas* las Gausianas cuando lo es para *cada*

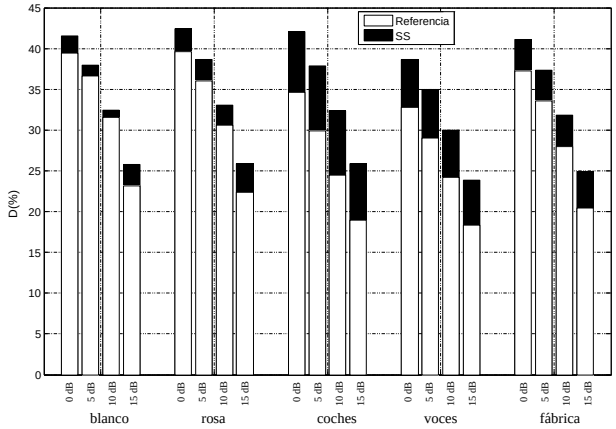


Figura 3.4: Porcentaje medio de la contribución de los *outliers* a la log-probabilidad acumulada en el grupo de test de la tarea RM1 para varios ruidos y SNRs. Las barras etiquetadas como *Referencia* se corresponden al cálculo de este término sobre los parámetros extraídos directamente a partir de la voz contaminada, mientras que las barras etiquetadas como *SS* se corresponden al cálculo de este porcentaje medio cuando se aplica sustracción espectral.

una de ellas, podemos concluir que el porcentaje D que estamos calculando es un subconjunto del porcentaje que calcularíamos con el método BD-HMM original.

La Figura 3.4 muestra el valor medio del porcentaje expresado mediante la ecuación (3.10) para el conjunto de frases de test de la base de datos RM1. Se muestran los resultados para los 5 tipos de ruidos y para las 4 SNRs consideradas en esta tarea. Además, presentamos los resultados aplicando SS (barras negras) y sin aplicar SS (barras blancas). Como vemos en la figura, la contribución de los *outliers* a la decisión final del reconocedor es muy significativa. En concreto, representan entre el 20 % y el 40 % de la log-probabilidad acumulada. Además, este valor medio del porcentaje aumenta sistemáticamente al aplicar SS. Parece evidente que, aunque los *outliers* no contienen ningún tipo de información lingüística, tienen un peso muy significativo en la decisión tomada en el reconocedor. Por lo tanto, el papel de BD-HMM para limitar esta contribución está totalmente justificado.

Tabla 3.2: WER (%) para voz limpia y los métodos a estudio para las bases de datos RM1, WSJ0 y Aurora-4 tasks.

	Referencia	**SS**	**BD-HMM**	**SSBD-HMM**
RM1	6.70 %	6.63 %	6.29 %	6.59 %
WSJ0	9.88 %	9.79 %	9.02 %	9.30 %
Aurora-4	9.38 %	9.79 %	8.84 %	9.25 %

3.4.4. Evaluación de nuestra propuesta en terminos de tasa de reconocimiento

La combinación de SS y BD-HMM, que denotamos de ahora en adelante como SSBD-HMM, se evaluó para las 4 tareas de reconocimiento previamente descritas. Los resultados se dan en términos de tasa de error en palabras (*Word Error Rate*, WER) y se consideran tres sistemas de referencia: en el primero no se aplica ninguna técnica robusta y la etiquetaremos como *Referencia*; en el segundo sistema se añade sustracción espectral al sistema de referencia y lo etiquetaremos como (*SS*); por último, con *BD-HMM* hacemos referencia al sistema de referencia cuando se aplica *bounded-distance HMM*.

Antes de presentar los resultados para cada tarea, mostramos en la Tabla 3.2 los resultados (WER) para voz limpia (los resultados para la base de datos *Spanish SDC-Aurora* no se muestran ya que no tenemos acceso a la voz limpia). Aunque BD-HMM proporciona ligeras mejoras en la tasa de reconocimiento, ningún método introduce cambios significativos respecto el sistema tomado como referencia.

Resultados y análisis adicional para la tarea RM1

Los resultados en términos de WER para la tarea RM1 y todas las técnicas a estudio, ruidos y SNRs se muestran en el Figura 3.5.

Como se observa en la figura, SSBD-HMM es la técnica que claramente obtiene los

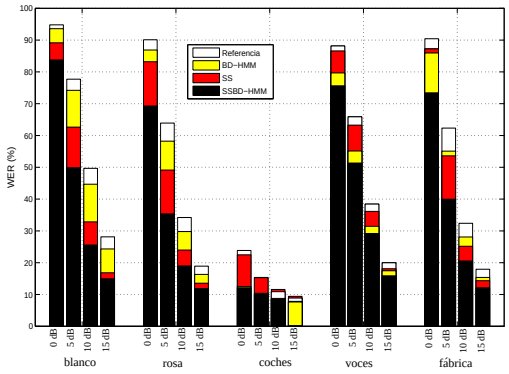


Figura 3.5: WER (%) para cada una de las técnicas a estudio para la tarea RM1.

mejores resultados con una excepción, el ruido de coches a 15 dB donde BD-HMM es ligeramente superior a SSBD-HMM. Además, las mejoras son estadísticamente significativas[2] para todos los ruidos con la excepción del ruido de coches, donde las prestaciones de SSBD-HMM y BD-HMM son similares.

Para los ruidos blanco, rosa y de fábrica SS aparece como una técnica más robusta que BD-HMM (con la excepción del ruido de fábrica a 0 dB). Para ruido de voces y de coches la situación es la inversa: BD-HMM mejora los resultados conseguidos por SS. El ruido de coches era el ruido para el que *acoustic-backing off* [de Veth et al., 2001b] consiguió los mejores resultados y nuestros resultados muestran exactamente la misma tendencia: BD-HMM consigue importantes mejoras y su combinación con SS no proporciona claras ventajas. Por otro lado, SS por si sólo tampoco consigue ninguna mejora reseñable e incluso se producen algunas pérdidas para las SNRs más altas. Sin embargo, estas pérdidas se compensan cuando combinamos SS con BD-HMM.

También debemos destacar la sinergia encontrada en la combinación de los dos métodos para la mayoría de los casos. Esta es evidente cuando comparamos las mejoras conseguidas con SSBD-HMM con la suma de las conseguidas por SS y BD-HMM. Esta comparación está ilustrada en la Figura 3.6 que muestra la reducción en

[2]La significación estadística ha sido evaluada calculando los intervalos de confianza para un nivel de confianza del 95 % [Weiss and Hasset, 1993].

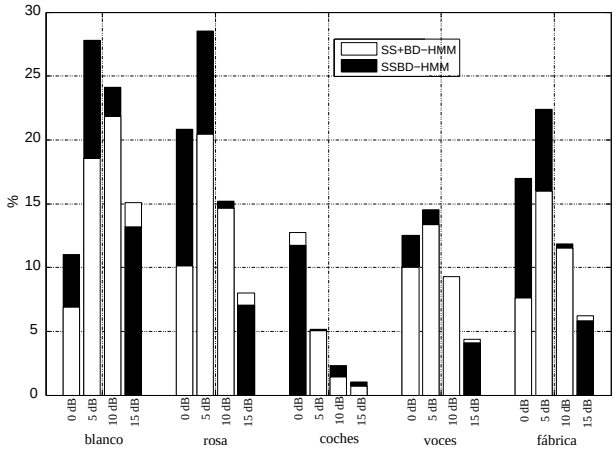


Figura 3.6: Comparación entre la reducción de la WER empleando SSBD-HMM y la suma de las reducciones conseguidas por SS y BD-HMM, etiquetado como *SS+BD-HMM*.

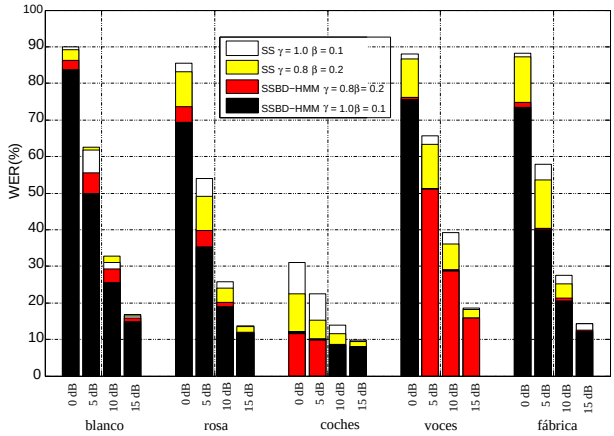


Figura 3.7: WER (%) para SS y SSBD-HMM para dos pares de parámetros, $\{(\gamma = 0{,}8; \beta = 0{,}2), (\gamma = 1{,}0; \beta = 0{,}1)\}$, la tarea RM1.

la tasa de palabras erróneas conseguida por SSBD-HMM y la suma de las conseguidas por SS y por BD-HMM, etiquetadas como SS+BD-HMM.

Como vemos en la figura, la sinergia es clara para SNR bajas (excepto para el ruido de coches). Este resultado concuerda con el hecho de que SS introduce un mayor número de *outliers* cuyo efecto es compensado mediante BD-HMM.

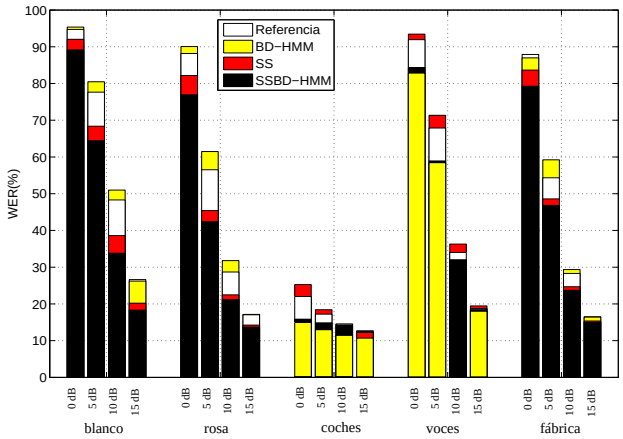


Figura 3.8: WER (%) conseguida por cada método con la tarea WSJ0.

Como indicamos en la Sección 3.3.3 los parámetros seleccionados para el método SS, γ y β, varían en función de que SS se aplique de forma aislada o en combinación con BD-HMM. En la Figura 3.7 presentamos los resultados (WER) para SS y SSBD-HMM y los dos pares de parámetros que se seleccionaron: $\{\gamma = 0{,}8;\ \beta = 0{,}2\}$ y $\{\gamma = 1{,}0;\ \beta = 0{,}1\}$.

En la mayoría de los casos, el par $\{\gamma = 0{,}8;\ \beta = 0{,}2\}$ es el óptimo cuando SS se aplica de forma aislada. Sin embargo, para SSBD-HMM el mejor par es generalmente $\{\gamma = 1{,}0;\ \beta = 0{,}1\}$. La razón a esta variación en el comportamiento la atribuimos al diferente número de *outliers* que genera SS en función del par empleado. Así, para el par $\{\gamma = 1{,}0;\ \beta = 0{,}1\}$ se genera un mayor número de *outliers* que para el otro par. Al aplicar BD-HMM, se reduce la influencia de estos *outliers* y se consiguen mejores resultados. Dicho de otro modo, el par $\{\gamma = 1{,}0;\ \beta = 0{,}1\}$ es más adecuado para compensar los parámetros pero, por otro lado, introduce un mayor número de *outliers*.

Resultados para las tareas WSJ0, Aurora-4 y Spanish SDC-Aurora

En esta subsección presentamos los resultados obtenidos para el resto de las tareas. En la Figura 3.8 presentamos los resultados para la tarea WSJ0. Para la mayoría

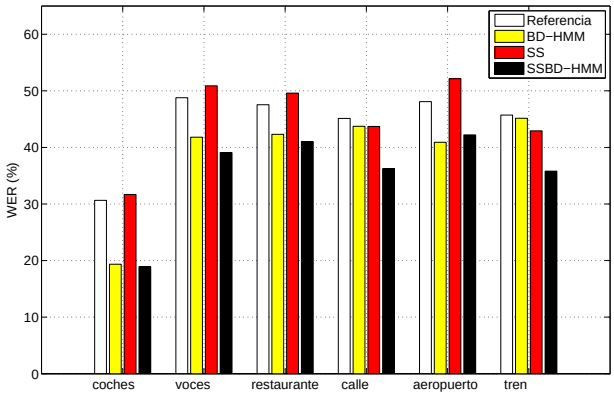


Figura 3.9: WER (%) conseguida para cada método con la tarea Aurora-4.

de los ruidos y SNRs la combinación SSBD-HMM consigue los mejores resultados. El estudio de los intervalos de confianza para estas tasas de error muestra que los resultados son estadísticamente significativos para SNRs bajas y medias. Encontramos dos excepciones a esta mejora en los ruidos de coches y de voces, en estos ruidos BD-HMM aplicado de forma aislada resulta ser la técnica más robusta mientras que SS no mejora los resultados. En cualquier caso, las prestaciones de SSBD-HMM y BD-HMM son generalmente comparables, es decir, BD-HMM compensa las pérdidas introducidas por SS. Debemos decir que, para estos dos ruidos, el par de parámetros que presentamos en la figura ($\{\gamma = 1{,}0;\ \beta = 0{,}1\}$) no fue el óptimo para SSBD-HMM y se obtuvieron mejores resultados con el par $\{\gamma = 0{,}8;\ \beta = 0{,}2\}$. Para el ruido de fábrica a SNRs medias los resultados avalan en mayor medida nuestra propuesta: BD-HMM no funciona y, en cambio, la combinación SSBD-HMM supera a SS. Idéntica situación observamos para el ruido blanco y rosa: SS es efectivo, BD-HMM no funciona y, sin embargo, la combinación claramente mejora los resultados obtenidos por SS haciendo evidente la apuntada sinergia.

Los resultados para la tarea Aurora-4 se muestran en la Figura 3.9. De nuevo, en la mayoría de los casos, la combinación de métodos consigue los mejores resultados. De hecho, SS generalmente no consigue ninguna mejora (con una excepción) y, sin

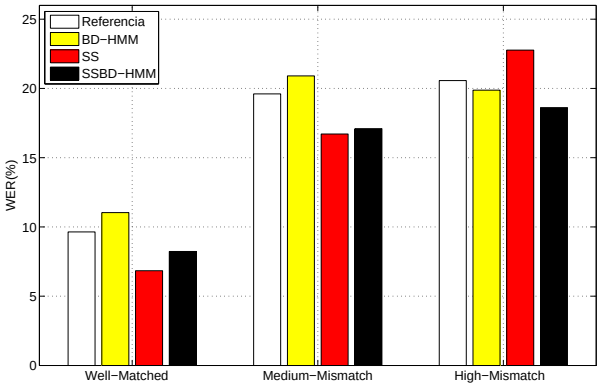


Figura 3.10: WER (%) conseguida para cada método con la tarea *Spanish SDC-Aurora task*.

embargo, la combinación SSBD-HMM mejora los resultados obtenidos por BD-HMM (de nuevo, con una excepción). Las mejoras son estadísticamente significativas para los ruidos de voces, calle y tren.

Para el ruido de coches y voces los resultados son similares a los obtenidos en la tarea WSJ0. Podemos hablar de una ligera mejora en los resultados conseguidos para el ruido de voces que probablemente se debe a la distinta SNR empleada para contaminar cada base de datos.

Por último, hemos evaluado nuestra propuesta con la tarea *Spanish SDC-Aurora* empleando un ruido que, en lugar de ser sumado artificialmente a la voz limpia, se captura directamente en un entorno real (Sección 3.4.2). En la Figura 3.10 se muestran los resultados para las tres condiciones estándares que se definen en la base de datos. De nuevo, la combinación SSBD-HMM es capaz de capturar lo mejor de cada método. La sinergia de ambos métodos se hace evidente para la condición más desfavorable (High-Mismatch, HM) donde BD-HMM consigue ligeras mejoras sobre los resultados de referencia, SS produce un empeoramiento notable del sistema y, en cambio, la combinación SSBD-HMM mejora significativamente los resultados conseguidos por el sistema tomado como referencia.

3.5. Conclusiones

En este capítulo proponemos la combinación de un método que denominamos *bounded-distance HMM* y sustracción espectral como una manera efectiva de paliar los efectos de los ruidos aditivos en el reconocimiento automático de habla.

BD-HMM se presenta como un método eficaz de reducir la influencia de los *outliers* en la toma de decisión llevada a cabo en el reconocedor. A lo largo del capítulo, hemos motivado el uso de BD-HMM cuantificando el peso que los *outliers* tienen en el cómputo de la log-probabilidad de las observaciones. También hemos estudiado las similitudes de BD-HMM con *acoustic-backing off* [de Veth et al., 2001a]. Los resultados experimentales nos permiten concluir que la combinación propuesta supera algunas de las limitaciones que presentaba tal método [de Veth et al., 2001b].

El alcance de BD-HMM es limitado ya que únicamente actúa sobre los *outliers*. Sin embargo, esta limitación se supera combinando BD-HMM y SS. De hecho, así lo muestran nuestros resultados que permiten concluir que SS es un aliado perfecto. Por un lado, mostramos cómo SS aumenta el número de *outliers* a la entrada del reconocedor pero que, gracias a BD-HMM, su efecto es limitado. De este modo, la combinación SSBD-HMM retiene lo mejor de cada método: la regeneración de todos los parámetros llevada a cabo por SS y la habilidad de limitar el efecto de los *outliers* (muchos de ellos introducidos por SS) por parte de BD-HMM. Por último, los resultados experimentales avalan el efecto de sinergia que permite concluir que la contribución de la combinación propuesta es mayor que la suma de las contribuciones de los métodos aplicados de forma aislada.

Capítulo 4

Decodificación con incertidumbre aplicada a los parámetros FF con sustracción espectral como método de regeneración de parámetros

4.1. Introducción

En el Capítulo 2 se hizo un breve repaso de las técnicas que tratan de compensar las distorsiones que afectan a la señal de voz; entre ellas, se mencionaron las basadas en "decodificación con incertidumbre". Tal y como vimos en aquel capítulo, estas técnicas incorporan en el proceso de decodificación la información referente a la incertidumbre que existe en las observaciones. En este capítulo aplicamos esta metodología a sistemas de reconocimiento que emplean la parametrización FF (*Frequency Filtered*) [Nadeu et al., 1995, Nadeu et al., 2001]. Esta parametrización, estudiada en la Sección 3.2, destaca por conseguir prestaciones similares a los coeficientes MFCC y permanecer en el dominio del log-espectro. Esta última característica hace que los métodos de decodificación con incertidumbre sean más fácilmente interpretables; lo

cual, como veremos, nos permite relacionar estos métodos con otros que ponderan la medida de similitud entre las observaciones y los modelos, de modo que se enfatizan los coeficientes del vector de parametrización más relevantes en presencia de ruido. Al permanecer en el dominio del log-espectro, estaremos hablando de métodos de ponderación espectral que enfatizan los picos del espectro frente a los valles.

Por otro lado, estos métodos se combinan con la técnica SSBD-HMM que presentamos en el capítulo anterior. Esta técnica estima los parámetros no contaminados por medio de sustracción espectral al mismo tiempo que reduce el impacto de los *outliers* en el reconocedor. Las técnicas de decodificación con incertidumbre que presentamos en este capítulo tienen en cuenta que, tras sustracción espectral, queda un cierto nivel de incertidumbre en las observaciones e incorporan esta información en el proceso de decodificación. De este modo, BD-HMM reduce el impacto de los *outliers* y, al mismo tiempo, sustracción espectral aplicado a un decodificador con incertidumbre compensa los efectos de las observaciones que, estando afectadas por el ruido, no están tan contaminadas como los *outliers*.

Los métodos propuestos se evalúan utilizando dos conocidas bases de datos contaminadas con ruidos aditivos. Los resultados ponen de relieve la conveniencia de incorporar información sobre la incertidumbre de las observaciones en el proceso de reconocimiento.

El resto de las secciones de este capítulo se organizan de la siguiente manera. En la Sección 4.2 detallamos nuestras propuestas: comenzamos estudiando el efecto que tienen los ruidos aditivos sobre la parametrización FF; tras este estudio, modelamos la incertidumbre que persiste en los parámetros FF y deducimos las nuevas reglas de decisión que gobiernan el reconocedor; a continuación, estas nuevas reglas de decisión se interpretan como métodos de ponderación espectral. En la Sección 4.3 se describen los experimentos y se muestran los resultados que evalúan las prestaciones de nuestras propuestas. Finalmente, en la Sección 4.4 presentamos las principales conclusiones deducidas a partir de nuestros experimentos.

4.2. Decodificación con incertidumbre aplicada a los parámetros FF con sustracción espectral como técnica de regeneración de parámetros.

En esta sección se presenta la metodología a seguir para aplicar la teoría de decodificación con incertidumbre sobre sistemas de reconocimiento que emplean los parámetros FF. Como método de regeneración de parámetros empleamos sustracción espectral.

En primer lugar estudiamos el efecto que tiene un ruido aditivo sobre los parámetros FF. A continuación, modelamos este efecto mediante una distribución de probabilidad que represente la incertidumbre y, finalmente, proponemos nuevos métodos de reconocimiento basados en decodificación con incertidumbre.

4.2.1. Efecto del ruido aditivo sobre los parámetros FF

Si bien en el capítulo anterior se empleó la potencia del espectro como salida de la etapa de análisis espectral, en este capítulo empleamos el módulo del espectro. Esto se debe a que, para analizar los efectos del ruido aditivo sobre la parametrización FF, hemos realizado una aproximación lineal del logaritmo del espectro; nosotros presumimos que esta aproximación lineal será más precisa cuando modelemos el logaritmo del módulo del espectro en lugar del logaritmo de la potencia del espectro.

El análisis que presentamos a continuación supone que el ruido es aditivo y no está correlacionado con la señal de voz original. Basándonos en esta hipótesis, asumimos que, en el dominio del módulo del espectro, voz y ruido también son aditivas. De este modo, nuestro análisis parte de esta hipótesis aditiva y determina el efecto del ruido sobre los parámetros FF.

Para comodidad del lector, reproducimos aquí la Figura 3.2 (ahora etiquetada como Figura 4.1) en la que presentábamos el diagrama de bloques que describe el cálculo de la parametrización FF. Como observamos en la figura, las energías en

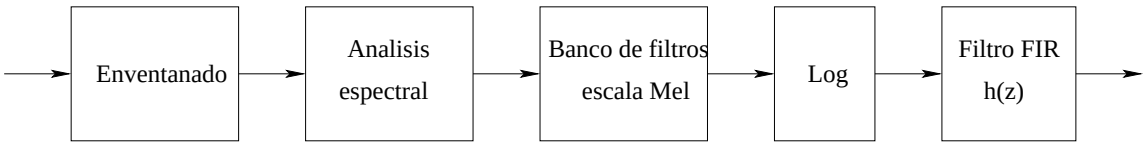


Figura 4.1: Diagrama de bloques de la parametrización FF

banda se calculan aplicando un banco de filtros al módulo del espectro. La salida de cada banco de filtros no es más que una combinación lineal de las componentes en frecuencia del módulo del espectro, por lo que la propiedad aditiva se conserva. De este modo, la relación entre las energías en banda contaminadas y las deducidas a partir de la voz limpia se puede escribir como:

$$\widehat{FBE}_k = FBE_k + n_k \tag{4.1}$$

donde k denota el índice de la banda correspondiente; $\widehat{FBE}_k$ la k-ésima energía en banda de la señal de voz contaminada; FBE_k la k-ésima energía en banda de la señal de voz limpia y, por último, n_k el ruido aditivo asociado a esa banda. Asumiremos que esta componente de ruido n_k es una variable aleatoria con media μ_{n_k} y varianza $\sigma^2_{n_k}$ y que las componentes de ruido en diferentes bandas de frecuencia están incorrelacionadas entre sí.

Debemos tener en cuenta que la sustracción espectral estima la media del módulo del espectro de ruido y la sustrae; como resultado, tendremos que la variable aleatoria que representa la componente de ruido en las energías en banda, n_k, tiene media cero ($\mu_{n_k} = 0$); es decir, finalmente n_k es una variable aleatoria con media cero y varianza $\sigma^2_{n_k}$.

Una vez calculadas las energías en banda procedemos a calcular las log-energías en banda:

$$\widehat{LFB}_k = \log(\widehat{FBE}_k) = \log(FBE_k + n_k). \tag{4.2}$$

Empleando la expansión de Taylor de primer orden del logaritmo alrededor de un cierto punto a llegamos a la siguiente expresión:

$$\widehat{LFB}_k \approx \log(a) + \frac{FBE_k}{a} - 1 + \frac{n_k}{a}. \tag{4.3}$$

Realizando la misma expansión alrededor del mismo punto de las log-energías en banda obtenidas a partir de la voz limpia obtenemos

$$LFB_k \approx \log(a) + \frac{FBE_k}{a} - 1. \tag{4.4}$$

Combinando las ecuaciones (4.3) y (4.4) obtenemos la expresión que relaciona las log-energías en banda contaminadas con las extraídas a partir de la voz limpia:

$$\widehat{LFB}_k \approx LFB_k + \frac{n_k}{a}. \tag{4.5}$$

Para que esta aproximación sea precisa el punto a debe estar próximo tanto a $\widehat{FBE}_k$ como a FBE_k. Como el primero es simplemente la estimación del segundo, que es desconocido, elegiremos como punto de aproximación $a = \widehat{FBE}_k = FBE_k + n_k$. Es interesante destacar que la cantidad de ruido presente en las log-energías en banda es inversamente proporcional al punto a, es decir, a FBE_k; pudiendo concluir que las bandas con energías mayores, es decir, los picos del espectro, se ven menos afectadas por el ruido que las bandas con energías bajas, comúnmente llamadas valles del espectro. Esto se debe a la función logaritmo: para energías altas, donde la derivada del logaritmo es baja, el oído es menos sensible a cambios en la presión acústica que a energías bajas, donde la derivada del logaritmo es mucho mayor.

Una vez deducida la relación entre las log-energías en banda contaminadas con las limpias, es sencillo determinar la relación de los parámetros FF contaminados con los limpios. Así, el k-ésimo coeficiente contaminado de la parametrización FF viene dado por:

$$\widehat{FF}_k = \widehat{LFB}_{k+1} - \widehat{LFB}_{k-1} \approx LFB_{k+1} + \frac{n_{k+1}}{a} - LFB_{k-1} - \frac{n_{k-1}}{b}, \tag{4.6}$$

donde $a = FBE_{k+1}+n_{k+1}$ y $b = FBE_{k-1}+n_{k-1}$. Dado que $FF_k = LFB_{k+1}-LFB_{k-1}$, la ecuación (4.6) puede ser reescrita del siguiente modo:

$$\widehat{FF}_k \approx FF_k + \frac{n_{k+1}}{a} - \frac{n_{k-1}}{b} \tag{4.7}$$

esto es,

$$\widehat{FF}_k \approx FF_k + N_k, \tag{4.8}$$

donde hemos denotado

$$N_k = \frac{n_{k+1}}{a} - \frac{n_{k-1}}{b} \tag{4.9}$$

a la nueva variable aleatoria que determina la incertidumbre de los parámetros FF.

La media y la varianza de esta nueva variable aleatoria vienen dadas por las siguientes expresiones:

$$\mu_{N_k} = 0 \tag{4.10}$$

$$\sigma^2_{N_k} = \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2}, \tag{4.11}$$

donde hemos asumido que las componentes de ruido en las distintas bandas están incorrelacionadas entre sí.

Todas estas ecuaciones se refieren a los parámetros estáticos de las observaciones y nada se ha comentado hasta ahora sobre las componentes de ruido que afectan a los parámetros dinámicos. Al ser los parámetros dinámicos una simple combinación lineal de las características estáticas en distintos instantes de tiempo, esa componente de ruido se puede calcular directamente a partir de la ecuación (4.8). Asumiendo que las componentes de ruido en distintos instantes de tiempo son independientes entre sí, el cálculo de la media y la varianza del ruido que afecta a los parámetros dinámicos también se calcula de forma directa a partir de las ecuaciones (4.10) y (4.11). En el apéndice A hemos detallado estas ecuaciones.

En función de la distribución de probabilidad asumida para la variable aleatoria que caracteriza el ruido de las observaciones nos encontramos con métodos basados en decodificación con incertidumbre diferentes. En el siguiente apartado detallamos las distribuciones de probabilidad asumidas a lo largo de esta tesis.

4.2.2. Modelado de la incertidumbre de los parámetros FF

En la Sección 2.6 revisamos las bases teóricas sobre las que descansan los métodos basados en decodificación con incertidumbre. Vimos que el criterio que gobierna en este caso el proceso de decodificación se expresa mediante la ecuación (2.18) que

reescribimos aquí:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i) p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) d\mathbf{o_t} \right] p(\lambda_i) \tag{4.12}$$

Para aplicar este nuevo criterio debemos determinar el tipo de distribuciones que modelan la incertidumbre de las observaciones, $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$. En este capítulo consideramos dos tipos: la primera asume que la variable aleatoria N_k en la ecuación (4.9) sigue una distribución Gausiana cuya media y varianza vienen dadas por las ecs. (4.10) y (4.11); la segunda es una Uniforme en la que su rango de actuación está directamente relacionado con la varianza de N_k.

Distribución de probabilidad Gausiana para el ruido

El primer caso que hemos considerado es aquél en el que asumimos que la distribución de probabilidad que modela la componente de ruido de los parámetros FF (ecuación 4.8) es una Gausiana:

$$N_{t_k} \sim \mathcal{N}(\cdot; 0, \sigma^2_{N_{t_k}}) \tag{4.13}$$

donde hemos añadido el subíndice t para hacer explícito el hecho de que esa componente de ruido depende de la trama que consideremos.

A partir de esta hipótesis vamos a deducir la expresión que tiene la distribución que modela la incertidumbre, $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$. En el caso particular que estamos considerando, el vector $\mathbf{o_t}$ está compuesto por los parámetros FF extraídos a partir de la voz limpia y las componentes del vector $\hat{\mathbf{o}}_\mathbf{t}$ se corresponden con las estimaciones obtenidas mediante sustracción espectral. Es decir,

$$\mathbf{o_t} = \begin{bmatrix} FF_{t_1} \\ \cdots \\ FF_{t_k} \\ \cdots \\ FF_{t_N} \end{bmatrix} ; \quad \hat{\mathbf{o}}_\mathbf{t} = \begin{bmatrix} \widehat{FF}_{t_1} \\ \cdots \\ \widehat{FF}_{t_k} \\ \cdots \\ \widehat{FF}_{t_N} \end{bmatrix} \tag{4.14}$$

siendo N la dimensión del vector de entrada al reconocedor (nótese la adición del subíndice t referente al índice temporal de la trama). En realidad, en este vector debemos añadir tanto la log-energía como los parámetros dinámicos; sin embargo, añadir estos parámetros complicaría la notación sin aportar ventajas adicionales. En cualquier caso, nuestra implementación ha tenido en cuenta la incertidumbre existente en los parámetros dinámicos de primer y segundo orden.

Si asumimos que los coeficientes a distintas frecuencias están incorrelacionados entre sí podremos escribir:

$$p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) = \prod_{k=1}^{N} p(FF_{t_k}|\widehat{FF}_{t_k}) \tag{4.15}$$

y a partir de las ecuaciones (4.8) y (4.13) es inmediato obtener la expresión que tendría la distribución de probabilidad que modela la incertidumbre de cada componente en frecuencia,

$$p(FF_{t_k}|\widehat{FF}_{t_k}) = \mathcal{N}(FF_{t_k}; \widehat{FF}_{t_k}, \sigma^2_{N_{t_k}}) \tag{4.16}$$

Una vez conocida la expresión de $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$ es posible actualizar la ecuación (4.12) que describe el comportamiento del reconocedor basado en decodificación con incertidumbre. Para ello asumiremos la distribución de probabilidad habitual para caracterizar los estados en los sistemas de reconocimiento basados en HMMs, la mezcla de Gausianas. De este modo, el término $p(\mathbf{o_t}|x_t^i)$ en la ecuación (4.12) adopta la siguiente expresión:

$$p(\mathbf{o_t}|x_t^i) = \sum_{m=1}^{M} c_{x_t^i m} \mathcal{N}(\mathbf{o_t}; \boldsymbol{\mu}_{\boldsymbol{x_t^i m}}, \boldsymbol{\Sigma}_{\boldsymbol{x_t^i m}}) \tag{4.17}$$

siendo M el número de mezclas de Gausianas, y $\boldsymbol{\mu}_{\boldsymbol{x_t^i m}}$ y $\boldsymbol{\Sigma}_{\boldsymbol{x_t^i m}}$ el vector de medias y la matriz de covarianzas que representan a la mezcla m del estado x_t^i.

Si consideramos ahora el caso más habitual en el que la matriz de covarianzas es diagonal, la ecuación (4.17) quedaría como sigue:

$$p(\mathbf{o_t}|x_t^i) = \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) \tag{4.18}$$

donde $\mu_{x_t^i m_k}$ representan la k-ésima componente del vector de medias $\boldsymbol{\mu_{x_t^i m}}$ asociado a la mezcla m del estado x_t^i; y, por otro lado, $\sigma^2_{x_t^i m_k}$ representa la k-ésima componente de la diagonal de la matriz de covarianzas $\boldsymbol{\Sigma_{x_t^i m}}$ asociada a la mezcla m del estado x_t^i.

Usando las ecuaciones (4.16) y (4.18) podemos reescribir la ecuación (4.12),

$$\lambda = \arg \max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \Bigg\{ \prod_{k=1}^{N} \int_{FF_{t_k}} \mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) \mathcal{N}(FF_{t_k}; \widehat{FF}_{t_k}, \sigma^2_{N_{t_k}}) dFF_{t_k} \Bigg\} \right] \quad (4.19)$$

Teniendo en cuenta las siguientes equivalencias para la segunda Gausiana en el integrando,

$$\mathcal{N}(FF_{t_k}; \widehat{FF}_{t_k}, \sigma^2_{N_{t_k}}) = \mathcal{N}(FF_{t_k} - \widehat{FF}_{t_k}; 0, \sigma^2_{N_{t_k}}) = \mathcal{N}(\widehat{FF}_{t_k} - FF_{t_k}; 0, \sigma^2_{N_{t_k}}), \quad (4.20)$$

la integral se reduce a la evaluación de la convolución de dos Gausianas en el punto $\widehat{FF}_{t_k}$:

$$\begin{aligned} &\int_{FF_{t_k}} \mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) \mathcal{N}(FF_{t_k}; \widehat{FF}_{t_k}, \sigma^2_{N_{t_k}}) dFF_{t_k} = \\ &= \Big[\mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) * \mathcal{N}(FF_{t_k}; 0, \sigma^2_{N_{t_k}}) \Big]_{FF_{t_k} = \widehat{FF}_{t_k}} \end{aligned} \quad (4.21)$$

donde el símbolo $*$ denota el operador convolución. La convolución de dos señales Gausianas da como resultado otra cuya varianza es igual a la suma de las varianzas de cada Gausiana. De este modo, la ecuación (4.19) puede ser reescrita como:

$$\lambda = \arg \max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \mathcal{N}(\widehat{FF}_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k} + \sigma^2_{N_{t_k}}) \right] \quad (4.22)$$

Así vemos como el criterio elegido para la decodificación con incertidumbre consiste simplemente en aumentar la varianza de las Gausianas que modelan los estados del reconocedor. Este aumento de varianza depende de la muestra temporal actual

ya que el término de ruido $\sigma^2_{N_{t_k}}$ varía con el tiempo y es función de las energías en banda.

La expresión (4.22) expresa la regla de decisión que adoptamos cuando aplicamos la decodificación con incertidumbre modelada mediante una distribución Gausiana. Si comparamos con el criterio que adoptaría un reconocedor convencional observamos que la única diferencia estriba en que la varianza de los modelos ha sido adaptada a los parámetros contaminados. Así lo vemos al calcular la media y la varianza de los parámetros FF contaminados en la ec. (4.8), la media sería la misma que la de los parámetros sin contaminar,

$$\mu_{\widehat{FF}_{t_k}} = E\{\widehat{FF}_{t_k}\} = E\{FF_{t_k}\} = \mu_{x_t^i m_k}, \tag{4.23}$$

pero la varianza necesita ser adaptada para representar correctamente los parámetros contaminados,

$$\sigma^2_{\widehat{FF}_{t_k}} = E\{(\widehat{FF}_{t_k} - \mu_{x_t^i m_k})^2)\} = \sigma^2_{x_t^i m_k} + \sigma^2_{N_{t_k}}. \tag{4.24}$$

Interpretación del modelo Gausiano de la incertidumbre como un método de ponderación espectral

Interpretar este método como un cambio de varianza nos permite a su vez hacer una nueva interpretación basada en métodos de ponderación del espectro [Vicente-Peña et al., 2006a]. Para ello partimos de nuevo de la ecuación (2.14) que describe el criterio por un reconocedor convencional y que reescribimos aquí considerando que, en lugar de maximizar la verosimilitud de las observaciones, maximizamos, de forma equivalente, la log verosimilitud de las observaciones:

$$\lambda = \arg\max_i \left(\log(a_{x_0^i x_1^i}) + \sum_{t=1}^{T} \left[\log(a_{x_t^i x_{t+1}^i}) + \log(p(\mathbf{o_t}|x_t^i)) \right] + \log(p(\lambda_i)) \right) \tag{4.25}$$

Fijándonos ahora en el término de la ec. (4.25) que hace referencia a la probabilidad de emisión de los estados, $\log(p(\mathbf{o_t}|x_t^i))$, y suprimiendo los términos que hacen referencia a los índices temporales podemos escribir:

$$\log(p(\mathbf{o}|x_t^i = j)) = -\frac{1}{2}\left\{\sum_{k=1}^{N}\log(2\pi\sigma_{j_k}^2) + \sum_{k=1}^{N}\frac{(FF_k - \mu_{j_k})^2}{\sigma_{j_k}^2}\right\} \tag{4.26}$$

donde hemos considerado que la probabilidad de emisión viene caracterizada por una sola Gausiana cuya media y varianza para la componente k-ésima vienen dadas respectivamente por μ_{j_k} y $\sigma_{j_k}^2$ (en la ecuación el estado actual x_t^i se ha igualado a j para simplificar las expresiones). Por otro lado, FF_k hace referencia a la componente k-ésima del vector de observaciones $\mathbf{o}$ de la voz limpia. El hecho de haber considerado Gausianas simples no resta generalidad ya que el análisis que estamos presentando se extrapola fácilmente al caso de tener mezclas de Gausianas. Sin embargo, consideramos que incluir explícitamente la información relativa a la mezcla de Gausianas oscurece el análisis sin aportar ningún concepto relevante.

Si ahora consideramos el caso en el que a la entrada del reconocedor tenemos vectores contaminados, debemos adaptar la varianza tal y como indica la ecuación (4.24). Así, reescribimos la ecuación (4.26),

$$\log(p(\hat{\mathbf{o}}|x_t^i = j)) = -\frac{1}{2}\left\{\sum_{k=1}^{N}\log\left(2\pi\frac{\sigma_{j_k}^2 + \sigma_{N_k}^2}{\sigma_{j_k}^2}\sigma_{j_k}^2\right) + \sum_{k=1}^{N}\frac{\sigma_{j_k}^2}{\sigma_{j_k}^2 + \sigma_{N_k}^2}\frac{(\widehat{FF}_k - \mu_{j_k})^2}{\sigma_{j_k}^2}\right\} \tag{4.27}$$

Introduciendo la notación

$$w_{j_k} = \frac{\sigma_{j_k}^2}{\sigma_{j_k}^2 + \sigma_{N_k}^2} \tag{4.28}$$

reescribimos la ecuación (4.27) como

$$\log(p(\hat{\mathbf{o}}|x_t^i = j)) = -\frac{1}{2}\left\{\sum_{k=1}^{N}\log(2\pi\sigma_{j_k}^2) + \sum_{k=1}^{N}w_{j_k}\frac{(\widehat{FF}_k - \mu_{j_k})^2}{\sigma_{j_k}^2} - \sum_{k=1}^{N}\log w_{j_k}\right\} \tag{4.29}$$

Comparando esta ecuación con la correspondiente a los parámetros limpios (ec. (4.26)) dos diferencias parecen claras:

- El término

$$\sum_{k=1}^{N}\frac{(FF_k - \mu_{j_k})^2}{\sigma_{j_k}^2} \tag{4.30}$$

para los parámetros limpios pasa a ser

$$\sum_{k=1}^{N} w_{j_k} \frac{(\widehat{FF}_k - \mu_{j_k})^2}{\sigma_{j_k}^2} \tag{4.31}$$

con los parámetros contaminados. Este término nos es más que la distancia Euclídea normalizada y nos indica cómo de cerca o lejos están nuestras observaciones del modelo representado a través de la media μ_{j_k} y la varianza $\sigma_{j_k}^2$.

Podemos considerar los pesos de la ecuación (4.28) como una medida de la cantidad de ruido presente en las observaciones. Así, cuando la varianza de la componente de ruido en nuestros parámetros, $\sigma_{N_k}^2$, es baja encontraremos pesos cercanos a la unidad. Sin embargo, estos pesos serán próximos a cero cuando esta varianza sea significativa. Normalmente, los pesos con valores próximos a uno provendrán de zonas donde el log-espectro contiene altas energías y, por tanto, es de esperar que la ec. (4.31) esté dominada por los picos del espectro en lugar de por los valles. Es también interesante destacar que los pesos dependen de la varianza de los modelos y por tanto los modelos con mayores varianzas serán los menos sensibles ante este tipo de distorsiones.

- la segunda diferencia cosiste en la adición del término

$$-\sum_{k=1}^{N} \log w_{j_k}. \tag{4.32}$$

El problema surge cuando todos los pesos, w_{j_k}, son próximos a cero en la ecuación (4.31) provocando una distancia Euclídea normalizada próxima a cero. Esta distancia tan próxima a cero significaría que la observación actual está muy bien representada por el modelo considerado cuando, en realidad, lo que ocurre es que las observaciones están muy contaminadas. La finalidad del término en la ecuación (4.32) es justamente evitar este tipo de situaciones e introduce una penalización para los pesos bajos. Debemos destacar que este

termino tiende a cero cuando los pesos tienen el valor 1, esto es, cuando no tenemos ruido.

Por último, hemos considerado la posibilidad de que la Gausiana representada mediante la ecuación (4.13) no sea la más adecuada para modelar la incertidumbre de las observaciones y hemos optado por introducir un grado de libertad más en dicha ecuación. Éste consiste en escalar la desviación típica de la Gausiana que modela el ruido de nuestros parámetros mediante un factor δ:

$$N_{t_k} \sim \mathcal{N}(\cdot; 0, (\delta\sigma_{N_{t_k}})^2) \tag{4.33}$$

Distribución de probabilidad Uniforme para el ruido

La segunda distribución de probabilidad que consideramos para modelar la incertidumbre de las observaciones (N_k en la ecuación (4.7)) es la distribución uniforme:

$$N_{t_k} \sim \mathcal{U}(\cdot; \widehat{FF}_{t_k}^{inf}, \widehat{FF}_{t_k}^{sup}) \tag{4.34}$$

donde $\widehat{FF}_{t_k}^{sup}$ y $\widehat{FF}_{t_k}^{inf}$ establecen los límites superior e inferior de la distribución y vienen determinados por las siguientes ecuaciones:

$$\widehat{FF}_{t_k}^{inf} = \widehat{FF}_{t_k} - \delta\sigma_{N_{t_k}} \tag{4.35}$$

$$\widehat{FF}_{t_k}^{sup} = \widehat{FF}_{t_k} + \delta\sigma_{N_{t_k}} \tag{4.36}$$

siendo δ un parámetro que controla el rango de actuación de la distribución Uniforme y $\sigma_{N_{t_k}}$ la desviación típica de la componente de ruido de las observaciones. El parámetro δ persigue el mismo propósito que el usado en la ecuación (4.33) para la distribución Gausiana: añadir un grado de libertad a las ecuaciones para así encontrar la distribución que mejor modele la incertidumbre de las observaciones. Nótese que de nuevo hemos incluido el subíndice t que hace referencia a la dependencia temporal de las expresiones.

Al igual que sucede con la distribución Gausiana, la distribución Uniforme también conduce a expresiones del reconocedor sencillas. De este modo, si modelamos nuestra incertidumbre a través de una distribución Uniforme,

$$p(FF_{t_k}|\widehat{FF}_{t_k}) = \mathcal{U}(FF_{t_k}; \widehat{FF}^{inf}_{t_k}, \widehat{FF}^{sup}_{t_k}) \tag{4.37}$$

$$= \frac{1}{2\delta\sigma_{N_{t_k}}} \begin{cases} 1 \text{ si } (\widehat{FF}^{inf}_{t_k} < FF_{t_k} < \widehat{FF}^{sup}_{t_k}) \\ 0 \text{ resto} \end{cases}, \tag{4.38}$$

podemos reescribir la ec. (4.12) como:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \left\{ \int_{FF_{t_k}} \mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) \right.\right.$$
$$\left.\left. \mathcal{U}(FF_{t_k}; \widehat{FF}^{inf}_{t_k}, \widehat{FF}^{sup}_{t_k}) dFF_{t_k} \right\} \right] \tag{4.39}$$

Particularizando en está última ecuación la expresión de la distribución Uniforme llegamos a la siguiente expresión que gobierna el comportamiento del reconocedor:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \frac{1}{2\delta\sigma_{N_{t_k}}} \int_{\widehat{FF}^{inf}_{t_k}}^{\widehat{FF}^{sup}_{t_k}} \mathcal{N}(FF_{t_k}; \mu_{x_t^i m_k}, \sigma^2_{x_t^i m_k}) dFF_{t_k} \right]. \tag{4.40}$$

Llegados a este punto es interesante destacar de nuevo la similitud que existe entre estos métodos y la técnica conocida como marginalización dentro de los métodos basados en las características más fiables. Esta semejanza, que ya fue descrita en la Sección 2.6.2, se hace ahora todavía más evidente observando la ecuación (4.40). Si el intervalo de integración abarca todos los valores posibles en una componente, estaríamos eliminando su efecto en el reconocedor (la integral sería igual a 1 para cualquier modelo).

Por otro lado, la integral de la ecuación (4.40) simplemente representa la probabilidad de que una variable aleatoria, FF_{t_k}, que se distribuye siguiendo una distribución Gausiana de media $\mu_{x_t^i m_k}$ y varianza $\sigma^2_{x_t^i m_k}$, tome valores comprendidos entre $\widehat{FF}^{inf}_{t_k}$

y $\widehat{FF}_{t_k}^{sup}$. Esta probabilidad es equivalente a que la variable aleatoria,

$$\frac{FF_{t_k} - \mu_{x_t^i m_k}}{\sigma_{x_t^i m_k}}, \tag{4.41}$$

que se distribuye según una Gausiana de media 0 varianza 1, tome valores comprendidos dentro del siguiente intervalo:

$$\left[\frac{\widehat{FF}_k^{inf} - \mu_{x_t^i m_k}}{\sigma_{x_t^i m_k}}, \frac{\widehat{FF}_k^{sup} - \mu_{x_t^i m_k}}{\sigma_{x_t^i m_k}} \right]. \tag{4.42}$$

Si definimos la función $F(x)$,

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy, \tag{4.43}$$

que calcula la probabilidad de que una variable aleatoria que se distribuye según una Gausiana de media 0 y varianza 1 tome valores menores que x, podemos reescribir la ecuación (4.40):

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \frac{1}{2\delta\sigma_{N_{t_k}}} \left\{ F\left(\frac{\widehat{FF}_k^{sup} - \mu_{x_t^i m_k}}{\sigma_{x_t^i m_k}} \right) - F\left(\frac{\widehat{FF}_k^{inf} - \mu_{x_t^i m_k}}{\sigma_{x_t^i m_k}} \right) \right\} \right] \tag{4.44}$$

Los valores de la función $F(x)$ pueden ser calculados fuera del reconocedor. Así, se sustituye el calculo de las integrales por simples accesos a tablas.

4.3. Experimentos y resultados

4.3.1. Descripción del sistema de reconocimiento y de los experimentos

Los métodos propuestos a lo largo de este capítulo han sido validados experimentalmente con dos bases de datos: RM1 [NIST, 1992] y Aurora-4 [Hirsch, 2002a]. Para cada base de datos se ha diseñado un sistema de reconocimiento usando la

herramienta HTK [Young et al., 2002]. Los sistemas construidos para estas bases de datos son idénticos a los explicados en la Sección 3.4 con únicamente dos salvedades: en la extracción de los parámetros se emplea el módulo del espectro en lugar de la potencia del espectro; sustracción espectral no se emplea para procesar las frases de entrenamiento.

De este modo, se emplean vectores de parametrización de 39 coeficientes que están compuestos por 12 parámetros FF, la log-energía y sus coeficientes dinámicos de primer y segundo orden. La versión de sustracción espectral empleada en los experimentos también coincide con la formulada en el Capítulo 3 con la diferencia de que ahora operamos en el dominio del módulo del espectro. Así, la única diferencia que encontramos en la ecuación (3.8) que describe dicho método es que en lugar de emplear las potencias del espectro ($\widehat{P_X}$, $\widehat{P_S}$ y $\widehat{P_N}$) se utilizan estimaciones del módulo del espectro. Las constantes γ y β conocidas respectivamente como "factor de sobre estimación" (*"over-estimation factor"*) y "nivel de suelo" (*"spectrum flooring"*) tomaron un valor igual a $\gamma = 0{,}8$ y $\beta = 0{,}2$ en nuestros experimentos. Estos parámetros se determinaron a partir de un pequeño barrido que consideró los valores $\gamma = \{0{,}8; 1{,}0; 1{,}2\}$ y $\beta = \{0{,}1; 0{,}2\}$ sobre la base de datos RM1 y ruidos contaminados a una SNR igual a 5 dB.

El estimador propuesto por [Martin, 2001] se utiliza para determinar el módulo del espectro del ruido. [Martin, 2001] asume que cada componente de la potencia del espectro sigue una distribución Exponencial. Ahora bien, , nosotros en lugar de la potencia del espectro trabajamos sobre su módulo de modo que la aproximación lineal realizada sobre el logaritmo es más precisa. Se sabe [Papoulis and Pillai, 2002] que el operador raíz sobre una variable aleatoria exponencial nos devuelve otra variable aleatoria que sigue una distribución Rayleigh. La media y varianza de esta distribución Rayleigh está relacionada con la media de la distribución Exponencial

Tabla 4.1: WER (%) para voz limpia y las bases de datos RM1 y Aurora-4.

	Referencia
RM1	6.57 %
Aurora-4	8.8 %

a través de las siguientes ecuaciones:

$$\mu_{RAY} = \frac{\sqrt{\pi}}{2}\sqrt{\mu_{EXP}} \tag{4.45}$$

$$\sigma^2_{RAY} = \left(1 - \frac{\pi}{4}\right)\mu_{EXP} \tag{4.46}$$

siendo μ_{RAY} y σ^2_{RAY} la media y la varianza de la distribución Rayleigh mientras que, μ_{EXP}, representa la media de la variable aleatoria exponencial o, en nuestro caso, la media de la potencia del espectro. A partir de la media y varianza de las ecs. (4.45) y (4.46) es inmediato caracterizar las componentes de ruido en las energías en banda.

Con el objetivo de determinar sobre qué parámetros son efectivas nuestras técnicas, las hemos combinado con el método SSBD-HMM que estudiamos en el Capítulo 3. SSBD-HMM es capaz de eliminar de forma eficaz el efecto de las muestras que están altamente contaminadas (*outliers*). De este modo, si los métodos propuestos en este capítulo son capaces de compensar los parámetros que no están tan contaminados, la combinación de ambos debería introducir mejoras sobre la aplicación de cada técnica de forma aislada.

Por último, antes de presentar los resultados sobre voz contaminada, en la Tabla 4.1 mostramos, en términos de tasa de error por palabras (WER, *Word Error Rate*) los que se consiguen para voz limpia.

4.3.2. Resultados

En esta sección presentamos los resultados obtenidos con los métodos propuestos. En primer lugar, mostramos los resultados para la base de datos RM1 y a continua-

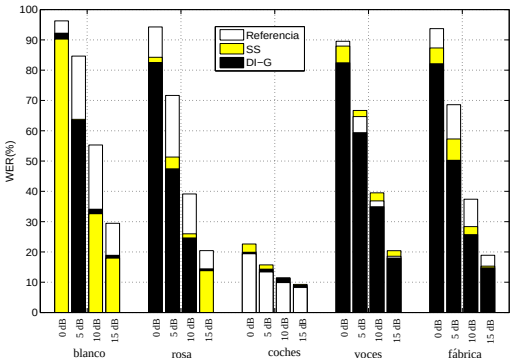


Figura 4.2: WER base de datos RM1: sistema de referencia (*Referencia*), aplicación de sustracción espectral (*SS*) y decodificación con incertidumbre empleando una distribución Gausiana para modelar la incertidumbre (*DI-G*).

ción lo hacemos para la base de datos Aurora-4.

Experimentos para la base de datos RM1

La Figura 4.2 muestra la tasa de error en palabras (WER) para la base de datos RM1 cuando empleamos una distribución Gausiana para modelar la incertidumbre de las observaciones (etiquetado como *DI-G* en la figura). En estos experimentos se utilizó el valor $\delta = 1{,}0$ en la ecuación (4.33). En la figura también mostramos los resultados de referencia obtenidos cuando no aplicamos ninguna técnica robusta (etiquetado como *Referencia*) y cuando aplicamos sustracción espectral (etiquetado como *SS*). Como vemos en la figura, la efectividad de los métodos varía en función de los ruidos a estudio. Así, para el ruido blanco parece que introducir la información de la incertidumbre de los parámetros no mejora las tasas de reconocimiento. Para el resto de ruidos, la decodificación con incertidumbre mejora los resultados obtenidos por SS. Debemos destacar los casos particulares del ruido de coches y el ruido de voces (para SNRs medias-altas) donde SS no introduce mejoras en el reconocimiento e incluso introduce algunas pérdidas. Sin embargo, la decodificación con incertidumbre es capaz de compensar esas pérdidas e incluso, para el ruido de voces, superar al

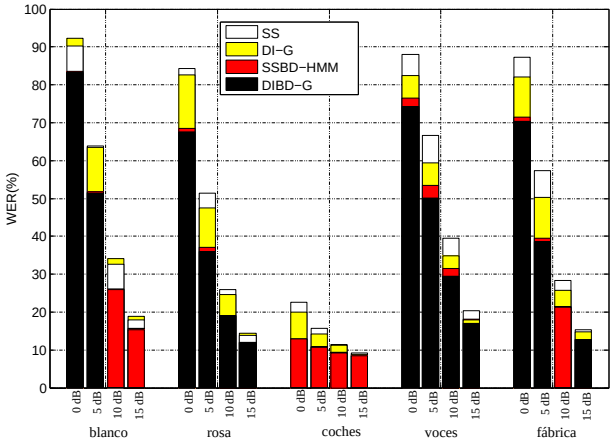


Figura 4.3: WER base de datos RM1: sutracción espectral (*SS*), decodificación con incertidumbre con distribución Gausiana (*DI-G*), SS combinado con BD-HMM (*SSBD-HMM*) y la combinación de la decodificación con incertidumbre con distribución Gausiana y SSBD-HMM (*DIBD-G*).

sistema tomado como referencia. Para el resto de las situaciones el comportamiento es el esperado, el sistema de referencia es superado por la aplicación de SS que a la vez es mejorado incorporando la información sobre la incertidumbre en las observaciones.

En la Figura 4.3 comparamos los resultados obtenidos por la decodificación con incertidumbre empleando una distribución Gausiana y los obtenidos mediante el método SSBD-HMM. Atendiendo a las tasas de error presentadas en la figura, SSBD-HMM se muestra como un método más robusto que el basado en decodificación con incertidumbre, por tanto, SSBD-HMM es más eficaz en el tratamiento de los parámetros altamente contaminados. Sin embargo, SSBD-HMM sólo tiene en cuenta este tipo de parámetros y pensamos que combinarlo con métodos basados en la decodificación con incertidumbre sería más efectivo ya que, estos últimos, podrían compensar parámetros que, estando distorsionados, no lo están de forma tan severa. En la misma Figura 4.3 mostramos la tasa de error en palabras conseguida para esta combinación (etiquetado como *DIBD-G*). Aunque las diferencias son generalmente poco significativas (con la excepción del ruido de voces donde las mejoras son es-

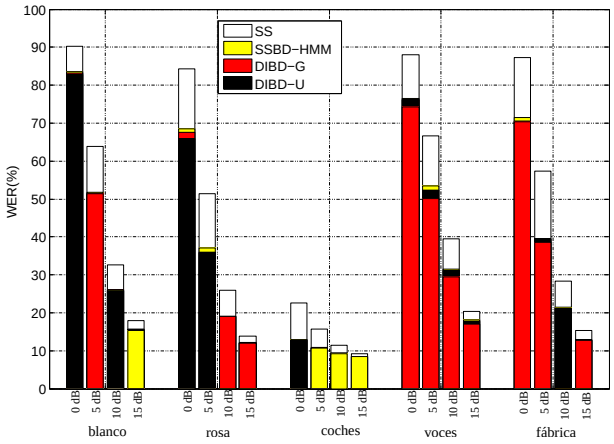


Figura 4.4: WER base de datos RM1. Comparación decodificación con incertidumbre con distribuciones de probabilidad Gausiana (*DIBD-G*) y Uniforme (*DIBD-U*) combinadas con SSBD-HMM. También se incluyen los resultados para SS y SSBD-HMM.

tadísticamente significativas), esta combinación consigue los mejores resultados para la mayor parte de las situaciones. Debemos decir que para evaluar la combinación se probaron varios valores del parámetro δ en la ecuación (4.33) que variaron entre 0,75 y 2,0. En el rango $\delta \in [0{,}75; 1{,}25]$ no se producen grandes variaciones en los resultados por lo que finalmente seleccionamos el valor $\delta = 1{,}0$ para los experimentos mostrados en la figura.

La Figura 4.4 muestra los resultados cuando se emplea una distribución Uniforme para modelar la incertidumbre de las observaciones. Los resultados se presentan combinados con SSBD-HMM (etiquetados como *DIBD-U*). Al igual que antes, se evaluaron valores de δ en las ecuaciones (4.35) y (4.36) en un rango comprendido entre 0,75 y 2,0. De nuevo, el sistema no es sensible a valores de δ situados alrededor de 1,0 y en la figura mostramos los resultados obtenidos para este valor. Como apreciamos en la figura, los resultados con una distribución Uniforme son generalmente equivalentes a los obtenidos con una distribución Gausiana.

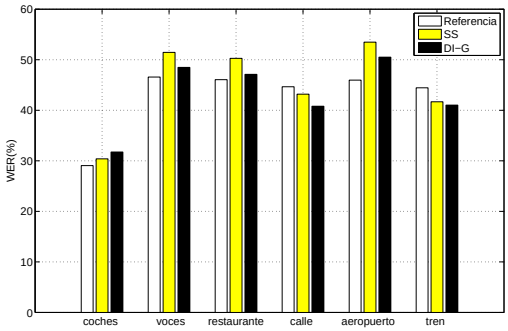


Figura 4.5: WER base de datos Aurora-4: decodificación con incertidumbre con distribución Gausiana (*DI-G*), sustracción espectral (*SS*) y el sistema sin aplicar ninguna técnica de robustez (*Referencia*).

Experimentos para la base de datos Aurora-4

En esta sección presentamos los resultados obtenidos para la base de datos Aurora-4. La secuencia de experimentos que describimos a continuación es la misma que la seguida para la base de datos RM1. En primer lugar, se evalúan las prestaciones que consiguen los métodos basados en la decodificación con incertidumbre de forma independiente. Por último, comparamos y combinamos estos métodos con SSBD-HMM. De este modo, SSBD-HMM limita el efecto de los *outliers* y los métodos basados en la decodificación con incertidumbre actuarían sobre el resto de los parámetros.

En la Figura 4.5 presentamos los resultados en términos de WER que la decodificación con incertidumbre consigue para esta base de datos. Excepto para el ruido de coches, la decodificación con incertidumbre supera las prestaciones de SS. Sin embargo, es el sistema de *Referencia* el que consigue los mejores resultados en la mayoría de los casos. Esto se debe a que SS no está funcionando de la forma esperada e introduce pérdidas para 4 de los 6 ruidos a estudio. Como quedó patente en el capítulo anterior, podemos paliar estas pérdidas que introduce SS combinándolo con BD-HMM. Así lo hacemos para el resto de experimentos.

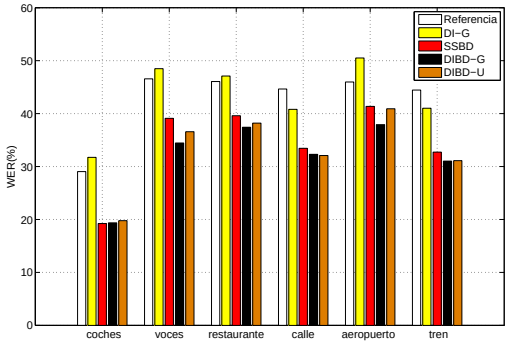


Figura 4.6: WER base de datos Aurora-4. Comparación de la decodificación con incertidumbre con SSBD-HMM.

En la Figura 4.6 comparamos las técnicas de decodificación con incertidumbre con el método SSBD-HMM. SSBD-HMM es claramente superior a los métodos basados en decodificación con incertidumbre por lo que también mostramos los resultados de la combinación de estos métodos con SSBD-HMM (modelando la incertidumbre o bien con una distribución Gausiana o bien con una Uniforme). A partir de los resultados podemos concluir que introducir información sobre la incertidumbre de las observaciones es adecuado para la mayoría de los ruidos. Cuando no lo es (ruido de coches), tampoco se pierden prestaciones. Modelar la distribución de probabilidad mediante una distribución Gausiana parece ser más conveniente para este tipo de ruidos. De nuevo, en estos experimentos hemos incorporado una constante de diseño de modo que se modifica el rango de actuación de las distribuciones de probabilidad. Así, para estos tipos de ruido hemos elegido una constante $\delta = 1{,}75$ para la distribución Gausiana (ver ec. (4.33)) y un valor $\delta = 1{,}25$ para la distribución Uniforme (ver ec. (4.35) y (4.36)). Estos valores fueron elegidos tras un barrido de valores entre 0,75 y 2,0 con pasos de 0,25. En general, las tasas de reconocimiento presentan una variación suave respecto a este parámetro. Para la distribución de probabilidad Gausiana (combinada con SSBD-HMM) los decrementos relativos de la WER respecto SSBD-HMM se sitúan en un 12 % para el ruido de voces, un 8 % para el ruido

de aeropuerto, alrededor de un 5 % para los ruidos tren y restaurante y, finalmente, un 3 % para el ruido calle. Para los ruidos de voces y aeropuerto los resultados son estadísticamente significativos.

4.4. Conclusiones

Los métodos propuestos en este capítulo parten de un sistema que reduce la cantidad de ruido presente en las observaciones mediante sustracción espectral (SS) e incorporan en el reconocedor la incertidumbre que todavía existe en los parámetros compensados. Nuestro trabajo se centra en sistemas que utilizan como *front-end* la parametrización FF, esta parametrización obtiene resultados tan buenos como la parametrización MFCC y presenta la ventaja adicional de permanecer en el dominio del log-espectro. Por ello, las ecuaciones que describen el efecto de los ruidos aditivos sobre esta parametrización son sencillas lo que facilita el modelado de su incertidumbre. Para dicho modelado se utilizan distribuciones Gausianas y Uniformes y, para cada distribución, se deducen las nuevas reglas de decisión que gobiernan el proceso de reconocimiento. Paralelamente, para el modelado Gausiano, la expresión de la nueva regla de decisión permite interpretar los métodos basados en decodificación con incertidumbre como técnicas de ponderado espectral.

Nuestros primeros resultados comparan estos métodos con SS y muestran la conveniencia de incorporar información de la incertidumbre de las observaciones en el proceso de reconocimiento.

Además, estos métodos basados en decodificación con incertidumbre se combinaron con la técnica SSBD-HMM que resulta ser muy efectiva en el tratamiento de *outliers*. El método SSBD-HMM combina los métodos BD-HMM y SS: BD-HMM se encarga de mitigar el efecto de los *outliers* mientras que SS compensa el resto de observaciones. Sin embargo, tras SS queda un cierto nivel de incertidumbre en las observaciones que es conveniente tratar mediante métodos basados en decodificación con incertidumbre. Nuestros resultados muestran que la combinación propuesta

mejora en general las prestaciones conseguidas por la aplicación de SSBD-HMM de forma individual. El modelado a través de la distribución Gausiana fue la que obtuvo mejores resultados aunque, para determinados ruidos, fue conveniente ampliar su rango de actuación incrementando su varianza.

Capítulo 5

Filtrado paso-banda de la evolución temporal de los parámetros espectrales para reconocimiento robusto en comunicaciones inalámbricas

5.1. Introducción

En este capítulo estudiamos el efecto que tienen los sistemas actuales de transmisión de voz sobre los sistemas de reconocimiento automático de habla. Dichos sistemas han evolucionado desde la red telefónica pública conmutada (PSTN, *Public Switched Telephone Network*) hasta un gran abanico de posibilidades que incluyen las redes móviles, la voz sobre IP, las redes *Bluetooth*, las redes inalámbricas locales o incluso las compuestas por combinaciones de varias de estas redes.

Por otro lado, estos nuevos sistemas de transmisión han permitido la proliferación

de nuevos servicios. Es aquí donde los sistemas de diálogo, basados en sistemas de reconocimiento automático de habla, pueden encontrar un gran nicho de mercado. Sin embargo todo esto sólo se conseguirá si se mejoran los sistemas de reconocimiento cuando la voz a reconocer ha sido transmitida a través de un canal inalámbrico. Este tipo de canales de transmisión introduce pérdidas de tramas y errores de transmisión que influyen negativamente en las tasas de acierto de los reconocedores. En este capítulo proponemos filtrar el espectro de modulación para aumentar la robustez de dichos sistemas de reconocimiento.

En el Capítulo 2 (Sección 2.2.2) estudiamos otros métodos que, basados en el filtrado del espectro de modulación, incrementan la robustez de los sistemas ante distorsiones tales como el cambio de micrófono o incluso ante ruidos aditivos (p. ej. [Hermansky and Morgan, 1994, Hanson and Applebaum, 1993, Nadeu et al., 1997]). En este capítulo adaptamos aquellas ideas a las nuevas distorsiones que aparecen en los sistemas de transmisión inalámbricos.

En este capítulo mostramos, tanto conceptualmente como experimentalmente, que filtrar paso-banda la evolución temporal de los parámetros espectrales es útil para combatir la distorsión introducida por la codificación de la señal de voz y por los errores de transmisión. Concretamente, proponemos dos nuevos métodos de extracción de parámetros: el primero, que denominamos BPF-LP-MFCC, consiste en un filtrado paso-banda de la evolución temporal de los parámetros LP-MFCC (ver Sección 1.1.1); el segundo, que denominamos M-RASTA-PLP, es una modificación del filtro empleado en el método RASTA-PLP [Hermansky and Morgan, 1994] por uno con la sección paso-bajo más abrupta. Ambas propuestas mejoran de manera significativa los resultados obtenidos tanto por la parametrización LP-MFCC como por el método RASTA-PLP en presencia de errores de transmisión [Vicente-Peña et al., 2006b].

Además, para limitar el efecto de los *outliers* en el reconocedor, hemos combinado estas nuevas parametrizaciones con el método BD-HMM, descrito en el Capítulo 3. En la sección experimental de este capítulo, mostramos que esta combinación mejora aún más la tasa de reconocimiento de nuestro sistema.

El resto del capítulo se organiza del siguiente modo: en la Sección 5.2 introducimos el problema del reconocimiento de habla cuando la señal de voz se transmite a través de un sistema de comunicación inalámbrico; en la Sección 5.3 describimos la manera en la que hemos afrontado este problema, basada en el filtrado del espectro de modulación; en la Sección 5.4 describimos los sistemas de reconocimiento empleados para validar nuestra propuesta y los resultados que obtiene en relación con los obtenidos por otras parametrizaciones robustas conocidas; finalmente, en la Sección 5.5, presentamos las conclusiones que se deducen a partir de nuestros resultados.

5.2. RAH en sistemas de comunicaciones inalámbricos

En un sistema de comunicaciones inalámbrico podemos identificar al menos tres tipos de distorsiones que afectan a los reconocedores automáticos de habla (RAH):

- Ambiente acústico: aunque este tipo de distorsión no se refiere de forma exclusiva a los sistemas inalámbricos, lo hemos incluido aquí explícitamente para hacer un mayor énfasis en el hecho de que, gracias a la comunicación inalámbrica, el acceso a servicios se realiza en múltiples y variados entornos que típicamente introducen ruido en el sistema. Por tanto, aunque sea de forma indirecta, los sistemas inalámbricos justifican en mayor medida la búsqueda de sistemas de reconocimiento más robustos.

- Distorsión de codificación: el ancho de banda de los sistemas inalámbricos es un recurso caro y la aparición de nuevos servicios inalámbricos no ha hecho más que empeorar la situación del saturado espectro radio-eléctrico. Por lo tanto, con el objetivo de optimizar el número de canales de transmisión se han elaborado múltiples e ingeniosos protocolos para compartir el limitado ancho de banda. Como parte de esos esfuerzos, la codificación de la señal de voz a tasas binarias bajas y medias ha potenciado la implantación de este tipo de redes y

su proliferación en el mercado. Esta compresión tan agresiva de la señal de voz introduce distorsiones que afectan a los sistemas de reconocimiento automático de habla.

- Errores de transmisión: los canales radio-eléctricos son, de manera intrínseca, variables y, en cierta medida, poco fiables. Es por ello que los errores de transmisión son más frecuentes en los sistemas de transmisión inalámbricos que en los guiados por cable. Aunque los sistemas de transmisión introducen codificadores de canal con el objetivo de minimizar estos errores, no son capaces de eliminar por completo su efecto que, en particular, también afecta a los sistemas de reconocimiento de habla.

Para paliar el efecto de este tipo de distorsiones, encontramos en la literatura tres maneras diferentes de abordar el reconocimiento: reconocimiento de habla local, distribuido y remoto. En la siguiente sección resumiremos brevemente estas tres arquitecturas prestando especial atención a la manera en la que evitan las nuevas distorsiones propias de este tipo de sistemas.

5.2.1. Arquitecturas de los sistemas de reconocimiento de habla en entornos inalámbricos

La clasificación que presentamos en esta sección sobre las diferentes arquitecturas de reconocimiento fue establecida por [Digalakis et al., 1999] atendiendo a la distribución de tareas entre el sistema local (o cliente) y el sistema remoto (o servidor). Con el fin de establecer esta clasificación distinguimos dos tareas principales: la extracción de características (*front-end*) y el reconocimiento (*back-end*).

Sistemas de reconocimiento locales

Las dos tareas se realizan en el dispositivo local y es por ello que hablamos de reconocimiento local [Junqua, 2000]. Este método es sin duda el más eficaz para evitar la distorsión de codificación y los errores de transmisión puesto que no necesitamos

transmitir la señal de voz para abordar el reconocimiento. El dispositivo local envía únicamente el texto transcrito al servidor.

El principal inconveniente de este tipo de sistemas se debe a la limitada capacidad de procesamiento de los dispositivos locales. De este modo, estos sistemas se ven limitados a tareas de reconocimiento sencillas con vocabularios restringidos. Nosotros hemos evaluado tareas de reconocimiento algo más complejas, lo que nos hace considerar otro tipo de arquitecturas.

Sistemas de reconocimiento distribuidos

En esta arquitectura las tareas se distribuyen entre el dispositivo local y un sistema remoto. De este modo, la extracción de los parámetros de reconocimiento se lleva a cabo en el dispositivo local mientras que el reconocimiento, con mayor demanda computacional, se lleva a cabo en el servidor remoto.

La principal ventaja de esta aproximación reside en el reducido ancho de banda que necesitamos para transmitir los parámetros que se emplean para realizar el reconocimiento. Además, la extracción de los parámetros no requiere, en general, de una excesiva carga computacional, por lo que es accesible a la mayoría de los dispositivos portátiles. Además, estos parámetros se podrían proteger a través de códigos más robustos que los que habitualmente se emplean para la transmisión de la voz codificada.

Por otro lado, existen multitud de investigaciones que buscan parametrizaciones más adecuadas a los sistemas actuales de reconocimiento. Algunas de ellas demandan un mayor coste computacional (p. ej. [Chen et al., 2004]) que las tradicionales pero, sin embargo, introducen importantes mejoras cuando se introducen en el lado del servidor. Obviamente esta arquitectura necesita que tanto el dispositivo local como el remoto acuerden el tipo de parámetros empleados para el reconocimiento y, por ello, se han realizado importantes esfuerzos de estandarización. Uno de los primeros estándares [ETSI ES 201 108, 2003] produjo resultados pobres en entornos contaminados y, por este motivo, se ha propuesto recientemente un nuevo estándar (*Advan-*

ced Front-End [ETSI ES 202 050, 2004]). Incluso así, actualmente no hay ninguna parametrización que pueda considerarse óptima independientemente del entorno de aplicación. Además, tal y como argumenta [Kiss et al., 2003], la implantación de estos sistemas distribuidos acarrea costes, debido a cambios en las infraestructuras actuales, que deben ser justificados por una clara mejora en el servicio proporcionado al usuario.

Sistemas de reconocimiento remotos

En esta última arquitectura el dispositivo local no necesita realizar ningún procesado adicional y únicamente necesita transmitir la señal de voz como lo hace habitualmente. De este modo, en el servidor encontramos el sistema reconocedor completo. Nosotros nos hemos decantado por esta arquitectura por los siguientes motivos:

- El servidor puede elegir la parametrización que mejor se ajuste a cada aplicación particular e incluso actualizarla si así lo necesita.

- No impone restricciones adicionales a los dispositivos locales en cuanto a capacidad computacional.

- Esta arquitectura cumple con los estándares actuales y no necesita de más ancho de banda para realizar su transmisión.

- Es posible recuperar la señal original con la calidad que proporcione el codificador de voz empleado.

Por otro lado, encontramos principalmente dos problemas en esta arquitectura remota: la distorsión de codificación y los errores de transmisión. En la siguiente sección discutiremos con más detalle estos dos inconvenientes y describiremos algunas de las soluciones propuestas en la literatura.

5.2.2. Distorsiones en los sistemas de transmisión inalámbricos

Las principales distorsiones que producen los canales de transmisión inalámbricos son la distorsión de codificación y la debida a los errores de transmisión.

Diversos autores han hecho hincapié en la pérdida de prestaciones de los reconocedores debido a la distorsión de codificación [Euler and Zinke, 1994, Lilly and Paliwal, 1996, Hirsch, 2002b]. Sin embargo, también es conocido que estas pérdidas se reducen de forma significativa entrenando el reconocedor con el mismo codificador de voz que el usado en la fase de test. Esta reducción es posible debido a que, a diferencia de lo que ocurre con distorsiones de otro tipo, como son las que producen los ruidos aditivos, se pueden igualar las condiciones de entrenamiento y de test, ya que el conjunto de codificadores de voz es limitado y quedan identificados a través de los protocolos de comunicación.

Existen soluciones que parten del flujo de bits codificados para reducir el efecto de los errores de transmisión [Pelaez-Moreno et al., 2001, Kim et al., 2002, Gallardo-Antolin et al., 2005]. Estas soluciones se basan en extraer los parámetros necesarios para abordar el reconocimiento a partir del flujo binario que representa la señal de voz. De este modo, no se necesita decodificar la señal de voz como paso previo al reconocimiento y aprovechamos el hecho de que los codificadores de canal asignan más bits a la codificación de la envolvente espectral y, por tanto, su transmisión es más robusta que la del resto de las componentes de la señal de voz. El principal inconveniente de este tipo de soluciones radica en el hecho de necesitar acceso al flujo de bits.

Habida cuenta de que muchas veces las aplicaciones no tendrían acceso al flujo de bits, nos parece oportuno el desarrollo de métodos de extracción de características que, a partir de la voz decodificada, consigan prestaciones competitivas ante las nuevas distorsiones propias de los canales inalámbricos. A lo largo de este capítulo analizamos estas distorsiones y proponemos filtrar el espectro de modulación para paliar sus efectos.

5.3. Filtrado del espectro de modulación para reconocimiento robusto en entornos inalámbricos

Como ya comentamos brevemente en las secciones anteriores, nuestra propuesta consiste en realizar un filtrado paso-banda del espectro de modulación para reducir el efecto de los errores de transmisión sobre los reconocedores automáticos de habla.

En la Sección 2.2.2 realizamos una revisión de las principales propuestas basadas en el filtrado del espectro de modulación para conseguir sistemas más robustos. Así, vimos como este tipo de filtrados resulta efectivo ante distorsiones convolutivas, siendo particularmente relevante la sección paso-alto. En esta sección, justificamos la adición de una sección paso-bajo; de este modo, proponemos filtrar paso-banda la evolución temporal de los parámetros espectrales con el objetivo de conseguir parametrizaciones más robustas ante errores de transmisión.

Los errores de transmisión propios de las comunicaciones inalámbricas producen a la entrada del decodificador de voz o bien errores residuales, que superan el decodificador de canal, o bien tramas que han sido descartadas completamente y sobre las que se ha aplicado el pertinente mecanismo de recuperación frente a errores. En nuestra propuesta tratamos ambos tipos de distorsión por medio de filtros.

Los errores de transmisión producen cambios impredecibles en las características espectrales de la señal de voz y, por tanto, podrían afectar a todo el espectro de modulación. Es decir, los errores residuales añaden un cierto nivel de aleatoriedad a los parámetros espectrales que podemos considerar como variaciones ruidosas en su evolución temporal. Nuestra conjetura es que estas variaciones producen componentes espurias en todo el ancho de banda del espectro de modulación.

Con respecto a las tramas descartadas, asumimos que la repetición (casi exacta) llevada a cabo por el mecanismo de tratamiento de errores genera tanto bajas como altas frecuencias en el espectro de modulación. Las primeras se deben a la evolución suave, prácticamente constante, de los segmentos repetidos y, las últimas al cambio abrupto que se produce cuando el decodificador recibe una trama sin errores.

Las hipótesis previas apoyan la idea de emplear un filtro paso-banda del espectro de modulación para atender únicamente a aquellas frecuencias que, siendo relevantes para la discriminación de los sonidos, están poco afectadas por los errores de transmisión.

Con el propósito de validar nuestras hipótesis hemos realizado una estimación del ancho de banda del espectro de modulación de los parámetros MFCC extraídos a partir de la voz limpia y extraídos a partir de voz decodificada que ha sido sometida a una transmisión inalámbrica. En estos experimentos, hemos definido el ancho de banda como el rango de frecuencias en el que encontramos el 90 % de la energía de nuestros parámetros [1]. El canal empleado en estos experimentos presenta una tasa de error de bit (BER *Bit Error Rate*) igual a $5 \cdot 10^{-2}$. En la Figura 5.1 presentamos el histograma de los anchos de banda calculados sobre los 6 primeros coeficientes MFCC. Para calcular este histograma hemos enventanado la evolución temporal de cada coeficiente MFCC y, sobre cada ventana, hemos calculado su ancho de banda.

Como se observa en la Figura 5.1 el efecto de los errores de transmisión en el espectro de modulación depende del coeficiente. En particular, para los dos primeros coeficientes observamos un mayor número de ventanas cuyo ancho de banda ha disminuido (el histograma está ligeramente desplazado hacia la izquierda). Esto es indicativo de la aparición de nuevas componentes de baja frecuencia en el espectro de modulación ya que el mismo porcentaje de energía está ahora concentrado en un menor ancho de banda. Por el contrario, para los coeficientes altos la dinámica es la opuesta, observando un mayor número de ventanas con un ancho de banda mayor en presencia de errores de transmisión (el histograma está ligeramente desplazado hacia la derecha). Los coeficientes de orden superior, no mostrados en la figura, siguen una tendencia similar a la de estos últimos.

Por tanto y con el propósito de reducir el efecto de los errores de transmisión, proponemos filtrar paso-banda la trayectoria temporal de los parámetros espectrales

[1] La metodología exacta utilizada en la simulación de los canales de transmisión y en la estimación de estos anchos de banda se explica de manera extensa en la Sección 5.4

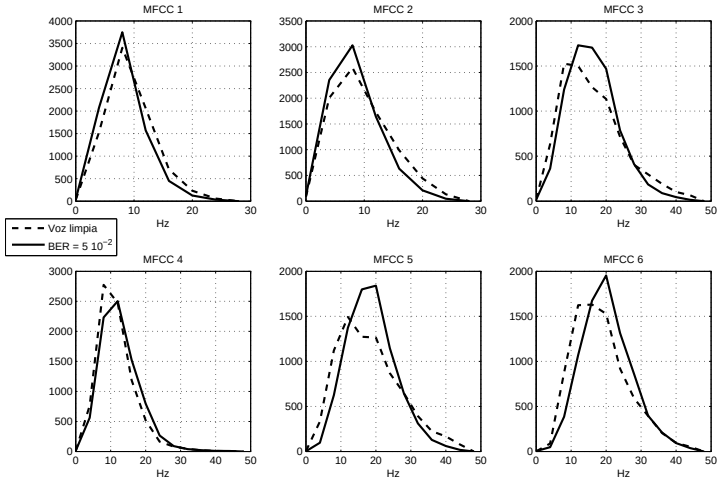


Figura 5.1: Histograma del ancho de banda del espectro de modulación para los seis primeros coeficientes MFCC extraídos a partir de voz limpia (trazo discontinuo) y a partir de voz con errores de transmisión (trazo continuo).

para, así, atenuar o eliminar estas componentes espurias de bajas o altas frecuencias que aparecen en el espectro de modulación. Atendiendo a los histogramas, parece que la solución óptima es la de utilizar un filtro paso-alto para los primeros coeficientes y un paso-bajo para los de mayor orden. Sin embargo, nosotros nos hemos inclinado por emplear el mismo filtro paso-banda para todos los coeficientes. Esta decisión ha sido adoptada por dos razones principalmente: es más sencillo de implementar y, por otro lado, debemos tener en cuenta que los histogramas muestran tan sólo una tendencia y lo más conveniente es eliminar todas aquellas componentes en frecuencia que son susceptibles de sufrir algún tipo de distorsión. Además, únicamente se eliminan aquellas frecuencias que no son importantes a la hora de diferenciar unos sonidos frente a otros (en presencia de alguna degradación, las componentes en frecuencia por encima de 10 Hz empeoran las prestaciones del reconocedor y las que están por debajo de 2 Hz no contribuyen a mejorarlas y, por tanto, podrían incluso degradar aún más los sistemas [Kanedera et al., 1998]).

A lo largo de este capítulo veremos cómo tanto el filtrado paso-bajo como el paso-

alto contribuyen de manera significativa a la mejora de las tasas de reconocimiento en presencia de errores de transmisión. En particular, proponemos replicar la sección paso-alto del conocido filtro RASTA [Hermansky and Morgan, 1994] y el diseño de una sección paso-bajo más abrupta como un buen compromiso entre preservar las frecuencias de modulación más relevantes y paliar los efectos de los errores de transmisión.

5.4. Experimentos y resultados

En esta sección describimos los experimentos que validan los métodos propuestos y mostramos sus resultados. En primer lugar se describe la base de datos y el entorno inalámbrico simulado. En segundo lugar, se especifican las principales características del sistema de reconocimiento empleado en los experimentos. En tercer lugar, se describe un experimento que estima el ancho de banda de los parámetros y que nos permite determinar el rango de frecuencias más relevante para cada coeficiente. A continuación se procede con los experimentos de filtrado, comenzamos cuantificando las mejoras que introduce un filtrado paso-bajo y seguimos con las producidas por un paso-banda, que constituye nuestra propuesta. Finalmente, se utiliza la técnica BD-HMM (descrita en el Capítulo 3) y se cuantifican sus mejoras en el entorno inalámbrico que estamos estudiando.

5.4.1. Descripción del sistema de reconocimiento y de los experimentos

Base de datos: Resource Management RM1 [NIST, 1992]

Para la realización de los experimentos se empleó la base de datos *Resource Management RM1* [NIST, 1992]. Esta base de datos coincide con la utilizada en los experimentos de los Capítulos 3 y 4 y fue descrita en la Sección 3.4.2. Al igual que en dichos capítulos, se realizan experimentos de habla continua independientes de

locutor empleando 3.990 frases para el entrenamiento y 1.200 para el reconocimiento. De nuevo, se emplea la versión submuestreada a 8 kHz.

Puesto que la base de datos se grabó en un entorno limpio, es posible estudiar los efectos de los errores de transmisión sin considerar otras interferencias que pudieran provocar otras fuentes de distorsión.

Modelo de canal inalámbrico

Con el objetivo de validar nuestra propuesta en condiciones realistas los experimentos se han realizado simulando un entorno GSM completo. Esta simulación incluye desde el modelo de canal hasta el proceso de codificación/decodificación de canal propuesto en GSM. El comportamiento del canal GSM se simula empleando un modelo híbrido que combina medidas experimentales (para modelar las zonas de sombra que se producen debido a la presencia de objetos tales como edificios en las zonas urbanas) y resultados teóricos (que modelan el desvanecimiento Rayleigh propio de las comunicaciones móviles). El codificador/decodificador de canal GSM se implementa siguiendo las especificaciones para canales de tráfico *half-rate* [ETSI Recommendation GSM 6.20, 1999]. Este incluye implementaciones del codificador de canal (tanto cíclico como convolucional) y los bloques necesarios para la construcción de las tramas TDMA GSM (reordenación, particionamiento, *interleaving* y formación de ráfagas). En [Gallardo-Antolin et al., 2005] podemos encontrar un mayor nivel de detalle sobre el simulador GSM usado en estos experimentos.

El modelo del canal inserta errores a ráfagas atendiendo a la tasa de error de bit (BER). El decodificador de canal es capaz de detectar y corregir algunos de estos errores o incluso sustituir tramas que están especialmente dañadas por una versión atenuada de la última trama recibida sin errores. De este modo, a la entrada del decodificador nos encontramos con dos tipos diferentes de errores: tramas que han sido borradas y tramas con errores residuales. El primero de ellos se mide a través de la tasa de tramas borradas (FER *Frame Erasure Rate*), que nos da el porcentaje de tramas que han sido reemplazadas por el mecanismo de tratamiento de errores; y

el segundo se caracteriza por la tasa residual de errores de bit (RBER *Residual Bit Error Rate*), que es el porcentaje de errores de transmisión que no son corregidos por el decodificador de canal.

Siguiendo este procedimiento, hemos diseñado cinco canales GSM *half-rate* que se corresponden con cinco condiciones del canal diferentes ($BER = 0$; 10^{-3}; 10^{-2}; $2{,}5 \cdot 10^{-2}$ y $5 \cdot 10^{-2}$). En la Tabla 5.1 presentamos los valores FER y RBER para cada uno de los canales simulados. Debemos decir que estos valores se han obtenido de manera experimental para la base de datos a estudio y no nos hemos limitado a su valor teórico.

Tabla 5.1: Características de los canales GSM *half-rate* usados en los experimentos. Mostramos las tasas BER, FER y RBER para cada canal

Canal	**BER**	**FER**	**RBER**
Canal 0	0	0 %	0 %
Canal 1	10^{-3}	0.015 %	0.0265 %
Canal 2	10^{-2}	0.479 %	0.2753 %
Canal 3	$2{,}5 \cdot 10^{-2}$	2.9296 %	0.8061 %
Canal 4	$5 \cdot 10^{-2}$	12.333 %	2.3222 %

La elección de estos canales se ha realizado atendiendo a la clasificación en calidades que realiza el estándar GSM [ETSI ETS 300 578, 1999]. Mostramos esta clasificación en la Tabla 5.2 que agrupa los canales en función de la BER que existe antes de la decodificación de canal.

En torno a la banda número cuatro encontramos la calidad media esperada en una comunicación GSM. Por ello hemos empleado canales cuya BER se sitúa alrededor de este rango. En concreto, hemos usado un canal en la tercera banda ($BER = 10^{-2}$), en la cuarta ($BER = 2{,}5 \cdot 10^{-2}$) y quinta ($BER = 5 \cdot 10^{-2}$). Además, también hemos usado dos canales que pertenecen a la banda que representa la mejor calidad: uno libre de errores ($BER = 0$, sólo tenemos distorsión de codificación) y otro con una

Tabla 5.2: Bandas de calidad en GSM

Banda de calidad	**BER**
0	$\text{BER} < 2 \cdot 10^{-3}$
1	$2 \cdot 10^{-3} < \text{BER} < 4 \cdot 10^{-3}$
2	$4 \cdot 10^{-3} < \text{BER} < 8 \cdot 10^{-3}$
3	$8 \cdot 10^{-3} < \text{BER} < 1{,}6 \cdot 10^{-2}$
4	$1{,}6 \cdot 10^{-2} < \text{BER} < 3{,}2 \cdot 10^{-2}$
5	$3{,}2 \cdot 10^{-2} < \text{BER} < 6{,}4 \cdot 10^{-2}$
6	$6{,}4 \cdot 10^{-2} < \text{BER} < 1{,}28 \cdot 10^{-1}$
7	$1{,}28 \cdot 10^{-1} < \text{BER}$

tasa de errores baja ($BER = 10^{-3}$).

5.4.2. Descripción del sistema base y resultados de referencia

Front-End

Para obtener los resultados de referencia a partir de los cuales comparar nuestras propuestas hemos empleado dos parametrizaciones bien conocidas: MFCC y LP-MFCC.

Estas parametrizaciones fueron estudiadas en la Sección 1.1.1 donde vimos cómo su única diferencia radicaba en la estimación del espectro. Así, la parametrización MFCC estima el espectro de la señal de voz mediante el cómputo directo de la transformada de Fourier mientras que, la parametrización LP-MFCC lo estima a través de un modelado todo polos de la señal de voz. En cuanto al orden del modelo todo polos se han probado experimentalmente (para voz limpia) los órdenes 8, 10, 12, 14 y 16 encontrando resultados similares para órdenes superiores a 10. Por tanto, para nuestros experimentos emplearemos modelos de orden 10.

En ambos casos hemos empleado ventanas Hamming de 25 ms, obteniendo 12 coe-

ficientes cada 10 ms. Las características estáticas así obtenidas se amplían añadiendo el coeficiente de log-energía. Finalmente, también añadimos los coeficientes de regresión de primer orden (ver Sección 1.1.1), de modo que la dimensión resultante del vector de parámetros es igual a 26.

Además, en los experimentos donde los parámetros son procesados mediante una etapa de filtrado, los parámetros dinámicos se calculan a partir de los parámetros filtrados. Los experimentos realizados por [Hanson and Applebaum, 1993] muestran mejores resultados cuando se calculan los parámetros dinámicos a partir de la secuencia filtrada. Estos experimentos se realizaron filtrando o bien las log-energías en banda o bien los coeficientes cepstrales que se obtienen a partir de un análisis PLP (*Perceptually-based linear prediction*), sin embargo, pensamos que empleando la parametrización MFCC o LP-MFCC se obtendrían conclusiones similares.

Back-End

Al igual que los sistemas descritos en el Capítulo 3, el *back-end* se basa en HMMs (*Hidden Markov Models*) y se utiliza la herramienta HTK [Young et al., 2002] para su implementación. De nuevo, se emplean modelos dependientes de contexto (*cross-word* trifonemas) que tienen 3 estados con 3 mezclas de Gausianas por estado. Los modelos se generan empleando o bien voz limpia (para obtener los resultados de referencia) o bien voz codificada, sin errores de transmisión. Los modelos que generamos a partir de la voz codificada son los que usamos para reconocer la voz que ha sufrido una transmisión inalámbrica. Finalmente, se emplea una bigramática como modelo de lenguaje.

Al reconocer los parámetros filtrados se emplean modelos entrenados con parámetros que también han sido filtrados. De este modo, evitamos que la etapa de filtrado introduzca desajustes adicionales entre las condiciones de entrenamiento y de test.

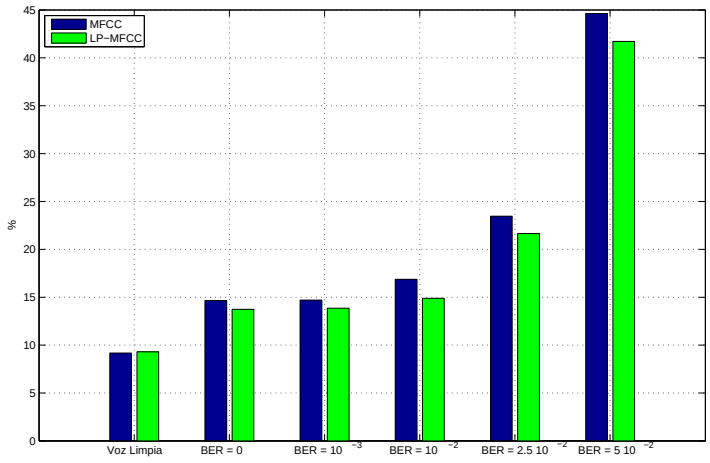


Figura 5.2: Resultados de referencia: WER(%) para las dos parametrizaciones (MFCC y LP-MFCC) y distintos canales GSM

Resultados de referencia

En la Figura 5.2 se muestran los resultados para las parametrizaciones MFCC y LP-MFCC. Los resultados se dan como la tasa de error por palabras (WER, *Word Error Rate*) que se produce para los datos de test de la base de datos. En la figura mostramos los resultados obtenidos para los distintos canales: voz limpia, voz con distorsión de codificación ($BER = 0$), y voz con distorsión de codificación y errores de transmisión ($BER = 10^{-3};\ 10^{-2};\ 2{,}5 \cdot 10^{-2}$ y $5 \cdot 10^{-2}$). Estos resultados serán los que emplearemos para futuras comparaciones.

A partir de estos experimentos extraemos nuestra primera conclusión: la parametrización MFCC obtiene prestaciones ligeramente superiores con voz limpia pero LP-MFCC es una parametrización más robusta ante distorsiones de codificación y errores de transmisión. Además, a medida que empeora el canal, más notables son las diferencias. El espectro suavizado obtenido a través del análisis LPC es probablemente el responsable de estas diferencias.

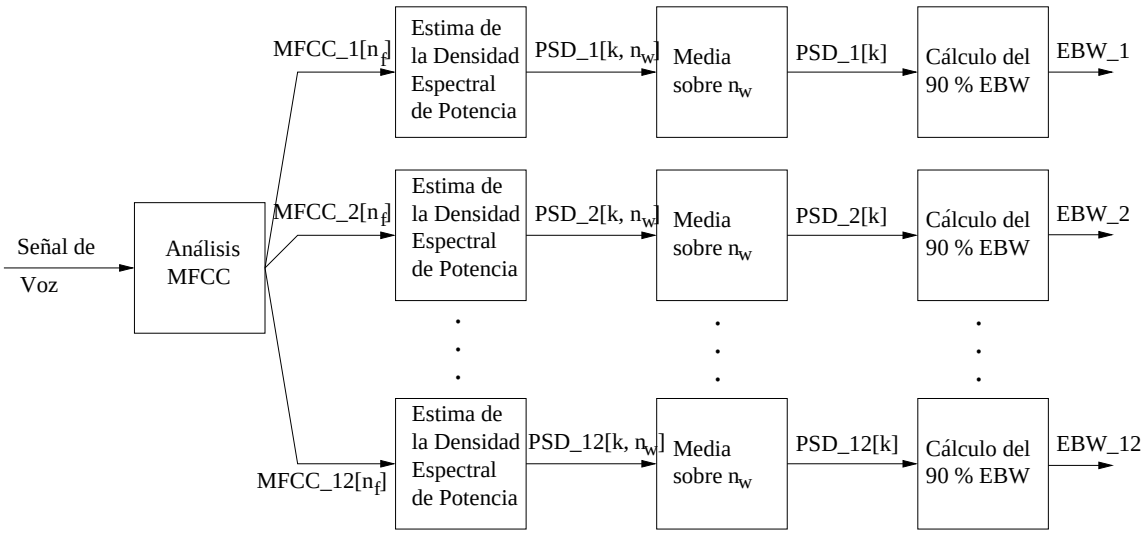


Figura 5.3: Proceso de estimación del ancho de banda efectivo para un porcentaje de energía del 90 % (90 % EBW)

.

5.4.3. Ancho de banda de los parámetros MFCC

Como paso previo al diseño de las secciones paso-bajo y paso-alto de los filtros que servirán para procesar la trayectoria temporal de los parámetros MFCC, hemos determinado experimentalmente las bandas del espectro de modulación que son más relevantes (asumimos que estas bandas coinciden con las bandas más relevantes de los parámetros LP-MFCC). En la Figura 5.3 hemos representado el esquema de bloques del proceso que nos permite determinar el ancho de banda para cada coeficiente. Este proceso consiste en las siguientes etapas [Peláez-Moreno et al., 2002]:

- En primer lugar se extraen 12 coeficientes MFCC ($MFCC\_i[n_f]$, donde $i = 1, 2 \ldots 12$ y n_f hace referencia al índice temporal) a partir de la voz limpia. En este caso usamos una tasa de tramas muy pequeña de modo que la estimación del ancho de banda de los parámetros no se ve afectada por posibles efectos de aliasing debido a un submuestreo. Esta tasa de tramas tan baja se emplea únicamente en estos experimentos y nunca para los experimentos de

reconocimiento de habla.

- A continuación analizamos la evolución temporal de cada coeficiente MFCC usando ventanas Hamming de gran duración. En concreto, usamos ventanas con una duración igual a 2 segundos con un solape entre ventanas del 50 %. Emplear una ventana de duración tan larga proporciona una resolución temporal muy pobre pero, sin embargo, la resolución en frecuencia es suficientemente buena como para poder estimar anchos de banda con precisiones alrededor de 1 Hz, algo necesario con los parámetros a estudio. A partir de cada ventana se calcula su densidad espectral de potencia. Estas señales están representadas en la Figura 5.3 con la notación $PSD\_i[k, n_w]$ $(i = 1, 2 \ldots 12)$ donde k representa el índice de la frecuencia de modulación y n_w representa el índice temporal de la ventana actual.

- Finalmente, calculamos la media de las densidades espectrales de potencia a lo largo del tiempo ($PSD\_i[k]$, $i = 1, 2 \ldots 12$). A partir de esta media calculamos el ancho de banda efectivo (EBW *Effective Bandwidth*) que definimos como el ancho de banda en el que queda concentrada una determinada porción de energía. Por ejemplo, el 90 % EBW se refiere al ancho de banda que contiene el 90 % de la energía de la señal actual (EBW_i, $i = 1, 2 \ldots 12$ en la Figura 5.3).

Un proceso similar se emplea para la estimación del ancho de banda de la log-energía.

En la Figura 5.4 representamos los anchos de banda estimados (90 % EBW) obtenidos para la log-energía y los 12 coeficientes MFCC. Como se observa en la figura, el ancho de banda de los parámetros aumenta con el orden del coeficiente comenzando alrededor de los 9 Hz para la log-energía y 11 Hz para el primer MFCC y yendo hasta los 32 Hz para el último coeficiente MFCC.

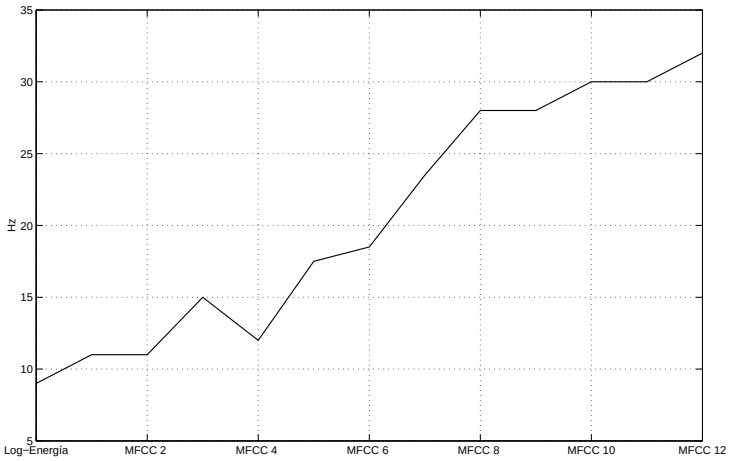


Figura 5.4: Ancho de banda efectivo para un porcentaje de energía del 90 % (90 % EBW) del coeficiente log-energía y los 12 coeficientes MFCCs

5.4.4. Filtrado paso-bajo

En la Sección 5.3 concluimos que un filtrado paso-banda del espectro de modulación de cada coeficiente mejoraría la robustez de los parámetros ante errores de transmisión. Hemos decidido diseñar este filtro en dos etapas: en primer lugar, se diseña un filtro paso-bajo de forma específica para tratar con los errores de transmisión y, en segundo lugar, a este filtro le añadimos una sección paso-alto. Procediendo de esta manera tendremos constancia de las aportaciones de cada sección a los resultados finales.

En esta sección abordamos el diseño del filtro paso-bajo mientras que, en la Sección 5.4.5, procedemos al diseño de la sección paso-alto.

Filtros FIR

Aunque los resultados previos acerca del ancho de banda de los parámetros parecen indicar que lo más conveniente es el empleo de diferentes filtros para cada coeficiente, después de algunos experimentos preliminares no encontramos claras ventajas

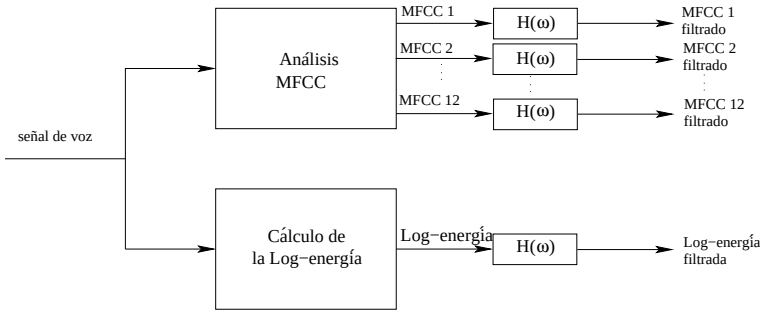


Figura 5.5: Filtrado de la parametrización MFCC

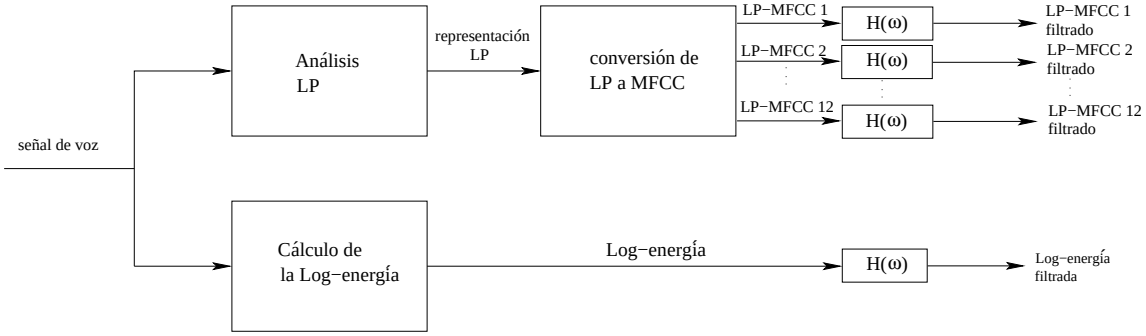


Figura 5.6: Filtrado de la parametrización LP-MFCC

al empleo de distintos filtros. De hecho, se encontraron resultados equivalentes a los obtenidos empleando el mismo filtro para todos los coeficientes, lo que no justifica el aumento de complejidad que conlleva la utilización de distintos filtros. De este modo empleamos un único filtro paso-bajo para todos los coeficientes y así evaluamos su efectividad cuando tanto los parámetros MFCC como los LP-MFCC son filtrados. En las Figuras 5.5 y 5.6 hemos representado el proceso de filtrado de estos coeficientes.

Aunque en la siguiente sección estudiaremos brevemente los filtros IIR, en ésta nos hemos centrado en los filtros FIR. Así, hemos estudiado como afectan sus principales características (orden y frecuencia de corte) a la tasa de reconocimiento conseguida para los distintos canales GSM. En concreto los experimentos se han realizado para las siguientes características de los filtros y condiciones:

- Orden: 10, 20 y 30

- Frecuencia de corte (Hz): 8, 10, 12, 18, 24 y 30.

- Ambiente: Voz limpia; $BER = 0;\ 10^{-2};\ 2{,}5 \cdot 10^{-2};\ 5 \cdot 10^{-2}$

concluyendo que el filtro más conveniente es aquel que tiene orden 20 y frecuencia de corte 12 Hz. Aunque un filtro de orden 20 parece ser demasiado largo teniendo en cuenta que potencialmente actúa sobre segmentos de la señal de voz de unos 200 ms, las mejoras en la tasa de reconocimiento (que mostramos más adelante) ante distorsiones de codificación y debidas a errores de transmisión hace que consideremos este orden como un buen compromiso entre la selectividad del filtro y lo que su respuesta al impulso se extiende en el tiempo.

Merece la pena mencionar que la frecuencia de corte así obtenida es cercana a la obtenida por otros autores como [Nadeu et al., 1997] o [Kanedera et al., 1998]. Si relacionamos esta frecuencia de corte con la estimación del ancho de banda de los parámetros (Figura 5.4) podemos observar que esta frecuencia de corte conserva prácticamente todo el espectro de modulación de los primeros cuatro coeficientes mientras que es bastante selectivo para el resto. En otras palabras, podemos considerar que este filtrado paso-bajo, además de eliminar las altas frecuencias del espectro de modulación, realiza un liftering de los coeficientes de orden superior.

En la Figura 5.7 mostramos los resultados para las dos secuencias de parámetros filtradas: LPF-MFCC (filtrado paso-bajo de los MFCC, *Low-Pass Filtered MFCC*) y LPF-LP-MFCC (filtrado paso-bajo de los LP-MFCC, *Low-Pass Filtered LPF-LP-MFCC*). Además, para facilitar la comparación, presentamos los resultados obtenidos con las secuencias de parámetros sin filtrar.

A partir de estos resultados, extraemos las siguientes conclusiones:

- Filtrar paso-bajo la secuencia temporal de los parámetros resulta efectivo para los reconocedores. Incluso en ausencia de distorsiones, el filtrado paso-bajo de los MFCC introduce algunas mejoras. Estas mejoras aumentan a medida que la condiciones de los canales empeoran.

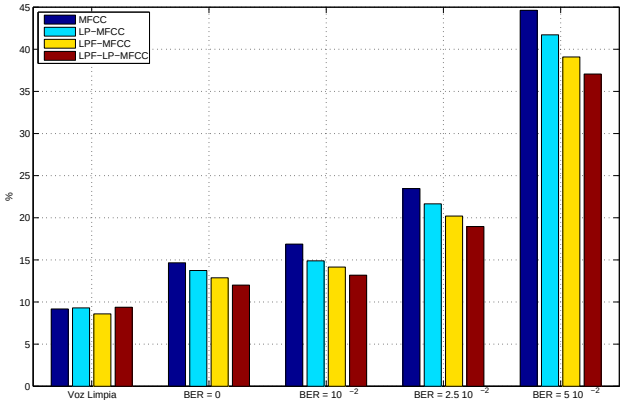


Figura 5.7: WER (%) para las secuencias filtradas LPF-MFCC y LPF-LP-MFCC. Además se incluyen los resultados para las secuencias sin filtrar, MFCC y LP-MFCC

- Cuando algún tipo de distorsión está presente, los mejores resultados se consiguen a través de la parametrización LPF-LP-MFCC. En concreto, la media de la reducción relativa de la tasa de error en palabras con respecto a la parametrización LP-MFCC es igual a 11.9 %.

Filtros IIR

Es posible diseñar un filtro IIR de menor orden que sea tan selectivo como el filtro FIR de orden 20 diseñado en la sección anterior. De este modo, aunque sólo sea desde el punto de vista computacional merece la pena evaluar los filtros IIR. Sin embargo, como el diseño de una implementación eficaz no es el objetivo de este trabajo, únicamente hemos realizado algunos experimentos preliminares para determinar si los filtros IIR podrían constituir una alternativa en el futuro.

En particular, se han probado filtros IIR de tipo Butterworth de orden 5 con frecuencia de corte 12 Hz. Se emplea este filtro con la parametrización LP-MFCC ya que, hasta ahora, es la que ha conseguido mejores resultados. Finalmente, los resultados obtenidos han sido sólo ligeramente inferiores a los obtenidos con los filtros FIR por lo que consideramos que este tipo de filtros IIR pueden ser una alternativa

más eficiente a los filtros FIR.

5.4.5. Filtrado paso-banda

En esta sección evaluamos la conveniencia de incluir una sección paso-alto para combatir las distorsiones debidas a la codificación de la voz y a los errores de transmisión.

Comenzamos evaluando el método RASTA-PLP ya que es una popular técnica que emplea un filtro paso-banda para procesar los parámetros en el dominio del log-espectro. Con el propósito de clarificar las diferencias entre el método RASTA-PLP y las técnicas de filtrado aquí propuestas, en la siguiente sección describimos brevemente este método.

RASTA-PLP (RelAtive SpecTrAl-Perceptually-based Linear Prediction)

En la Sección 2.2.2 repasamos brevemente el método RASTA-PLP encuadrándolo como una de las primeras técnicas basadas en el filtrado del espectro de modulación. En esta sección estudiamos un poco más a fondo este método comparándolo con las técnicas de filtrado propuestas en las secciones anteriores.

En la Figura 5.8 presentamos un esquema de bloques del método RASTA-PLP [Hermansky and Morgan, 1994]. Comparándolo con el esquema de bloques que ilustra el cómputo de la parametrización PLP (Figura 1.6) encontramos una nueva etapa de filtrado en el dominio del logaritmo del espectro que se lleva a cabo a través de los siguientes pasos:

- LOG (logaritmo): A través de este operador convertimos las energías en bandas al dominio del logaritmo del espectro. El logaritmo del espectro presenta la ventaja de que las distorsiones convolutivas se transforman en aditivas.

- Filtrado paso-banda de la evolución temporal de las log-energías en banda. El filtro que emplea RASTA-PLP presenta la siguiente función de transferencia:

$$H(z) = 0{,}1z^4 \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - \rho z^{-1}}. \tag{5.1}$$

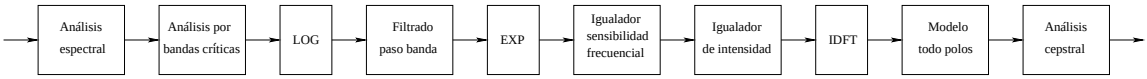


Figura 5.8: Análisis RASTA-PLP

Como vemos en la ecuación, este filtro tiene 4 ceros y un polo. Originalmente, los autores del método RASTA-PLP [Hermansky and Morgan, 1994] asignan un valor al polo igual a 0.98 pero encuentran conveniente realizar un pequeño barrido para adaptarse a cada tarea en particular.

- EXP (exponencial): Aplicamos el operador exponencial para así volver al dominio de las energías en banda y poder proceder con el resto de las etapas involucradas en el cálculo de la parametrización PLP.

En cuanto al orden del modelo todo polos empleado para el método RASTA-PLP, empleamos un orden igual a 12 tras evaluar distintos ordenes (8, 10, 12, 14 y 16) con voz limpia y distorsionada (distorsiones de codificación y debida a los errores de transmisión). Además, también se ha variado la posición del polo en el filtro RASTA-PLP (ρ en la ecuación 5.1). En concreto hemos evaluado valores del polo que van desde 0.5 hasta 0.98 para voz limpia y distorsionada. Se observa que a medida que el polo se aproxima a la unidad los resultados mejoran. Sin embargo, para los valores más altos las diferencias no son significativas y por ello hemos fijado la posición del polo, ρ, a un valor igual a 0.98.

La Figura 5.9 muestra la tasa de error en palabras para las parametrizaciones PLP, RASTA-PLP y LPF-LP-MFCC para distintas condiciones en los canales. El filtrado paso-banda que realiza el método RASTA-PLP mejora los resultados encontrados por PLP cuando existen distorsiones de codificación y errores de transmisión. Por otro lado, mientras que LPF-LP-MFCC mejora los resultados que consigue RASTA-PLP para los peores canales, aquellos con BER mayor que $2{,}5{\cdot}10^{-2}$, RASTA-PLP es la mejor solución para los canales con baja BER.

Estos resultados permiten extraer las siguientes conclusiones:

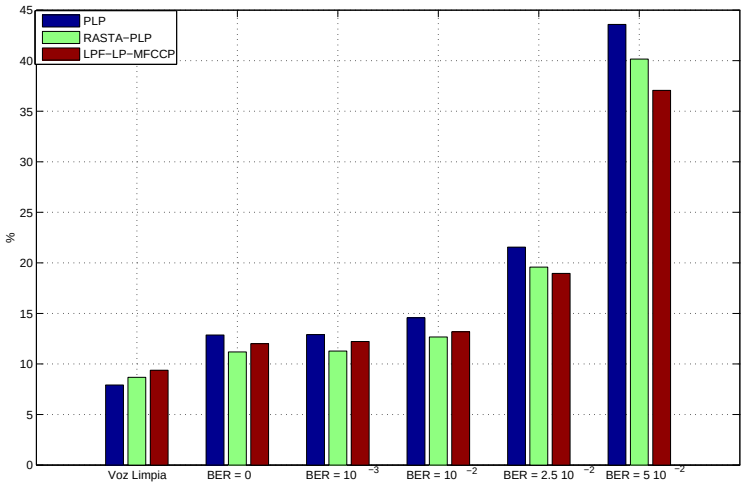


Figura 5.9: WER (%): Comparación entre las parametrizaciones PLP, RASTA-PLP y LPF-LP-MFCC

- RASTA-PLP es efectivo para combatir las distorsiones de codificación y las debidas a errores de transmisión
- La sección paso-bajo del filtro de RASTA no es tan selectiva como lo requieren los canales de alta BER.

Esta última conclusión nos lleva a proponer un nuevo método que sustituye la parte paso-bajo del filtro RASTA-PLP por otra con aspecto similar a la de la parametrización LPF-LP-MFCC.

Combinación de la sección paso-alto del filtro RASTA con una sección paso-bajo más selectiva

Se diseña un filtro paso-banda combinando la sección paso-alto del filtro RASTA-PLP (con $\rho = 0{,}98$) y la sección paso-bajo del filtro FIR propuesto para la parametrización LPF-LP-MFCC (FIR de orden 20 con frecuencia de corte igual a 12 Hz). El filtro paso-banda así diseñado se aproxima usando un filtro IIR de 20 ceros y 1

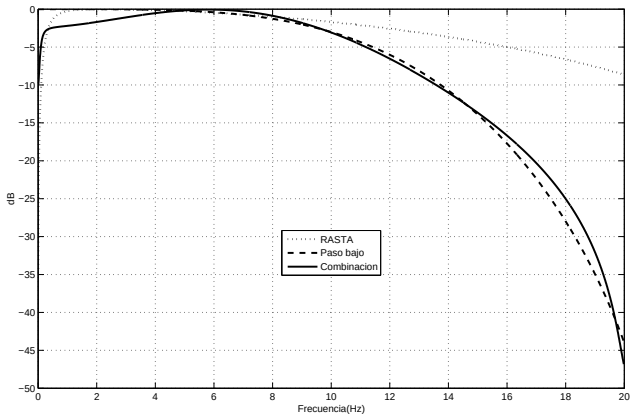


Figura 5.10: Módulo de la respuesta en frecuencia del filtro RASTA, el filtro FIR paso-bajo y el filtro paso-banda diseñado como la combinación de los dos anteriores

polo. En la Figura 5.10 hemos representado el módulo de la respuesta en frecuencia de este filtro (la fase ha sido elegida lineal).

Hemos evaluado el nuevo filtro sobre dos parametrizaciones: 1) filtrando paso-banda los parámetros LP-MFCC que pasamos a denotar como BPF-LP-MFCC (Band-Pass Filtering LP-MFCC); y 2) como alternativa al filtro paso-banda que se emplea en el método RASTA-PLP que denominamos M-RASTA-PLP (*Modified-RASTA-PLP*).

En la Figura 5.11 presentamos los resultados (tasa de error en palabras) para la parametrización BPF-LP-MFCC comparándola con la parametrización LPF- LP-MFCC para todos los canales considerados (también incluimos los resultados para la parametrización sin filtrar para que así sirva de referencia). Como puede observarse, BPF-LP-MFCC siempre consigue los mejores resultados con decrementos relativos (respecto LPF-LP-MFCC) que van desdel el 2 % para una BER igual a 0 hasta un 14 % para la BER igual a $5 \cdot 10^{-2}$ (para la voz limpia el decremento relativo es del 6 %). Por lo tanto, podemos concluir que la sección paso-alto del filtro también es útil para combatir las distorsiones que afectan a la voz en ambientes inalámbricos. Además, las ventajas de incluir la sección paso-alto se hacen más evidentes a medida

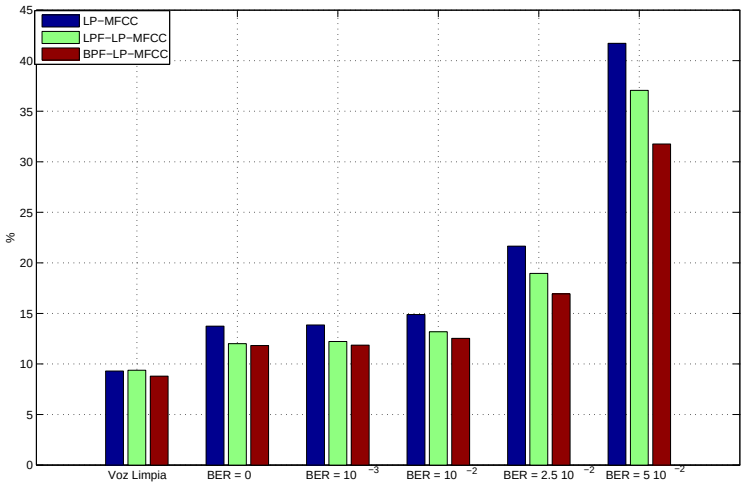


Figura 5.11: WER: Comparativa entre los métodos BPF-LP-MFCC y LPF-LP-MFCC. Los resultados para la parametrización LP-MFCC se muestras como referencia

que la calidad de los canales empeora. En particular, las mejoras son estadísticamente significativas para los canales con BERs iguales a $2{,}5 \cdot 10^{-2}$ y $5 \cdot 10^{-2}$.

En la Figura 5.12 presentamos los resultados obtenidos por la parametrización M-RASTA-PLP en relación con los logrados por el método RASTA-PLP (de nuevo, los resultados para la parametrización sin filtrar, PLP, se muestran como referencia). Bajo todas las condiciones a estudio, el método M-RASTA-PLP consigue mejores resultados que el método RASTA-PLP y las mejoras, debido a hacer la sección paso-bajo más selectiva, son mayores a medida que las condiciones del canal empeoran. En particular, los resultados son estadísticamente significativos para BERs iguales a $2{,}5 \cdot 10^{-2}$ y $5 \cdot 10^{-2}$.

Por último, en la Figura 5.13 hemos realizado una comparación entre los métodos sobre los que hemos encontrado los mejores resultados: BPF-LP-MFCC y M-RASTA-PLP. Aunque las diferencias no son estadísticamente significativas, la tendencia parece clara: BPF-LP-MFCC supera a M-RASTA-PLP en los canales con alta BER

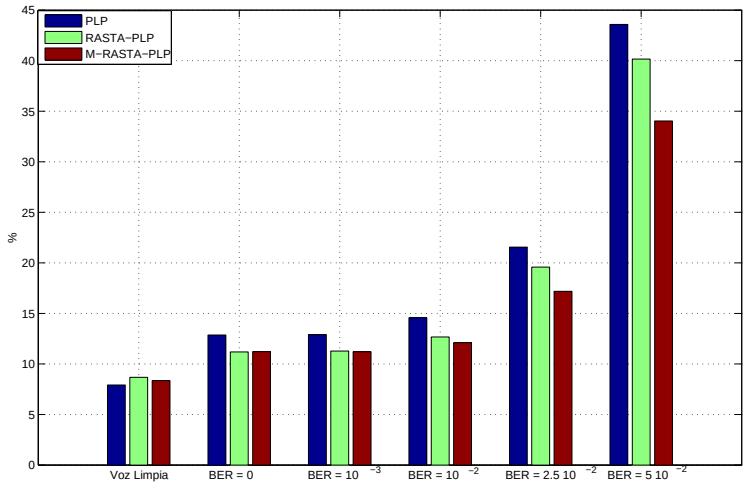


Figura 5.12: WER: Comparativa entre los métodos M-RASTA-PLP y RASTA-PLP. Los resultados para la parametrización PLP se muestran como referencia

mientras que M-RASTA-PLP es la mejor solución para los canales con baja BER. Pensamos que estos resultados se deben al lugar donde se hace el modelado todo polos del espectro de la señal de voz. En la parametrización PLP, el modelado se realiza en las últimas etapas mientras que, para la parametrización LP-MFCC, este modelado tiene lugar en las primeras etapas. Aunque deberemos confirmarlo con futuras investigaciones, nuestra primera intuición es que el suavizado debe realizarse en las primeras etapas cuando tratamos con canales con alta BER.

5.4.6. BD-HMM y comunicaciones inalámbricas

Por último, con el objetivo de limitar el efecto de los *outliers* en el reconocedor, hemos utilizado la técnica BD-HMM que explicamos en la Sección 3.2. La Figura 5.14 muestra los resultados cuando aplicamos esta técnica sobre las principales parametrizaciones que hemos estudiado en este capítulo. En dicha figura presentamos con una textura lisa los resultados empleando el reconocedor convencional y, con textura rayada, los obtenidos aplicando BD-HMM. A diferencia de lo ocurrido en los

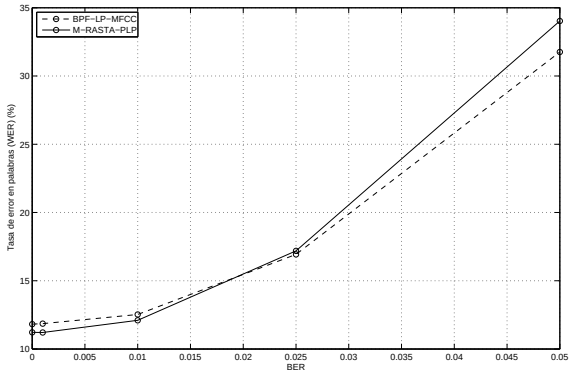


Figura 5.13: WER: Comparativa entre los dos mejores métodos estudiados: BPF-LP-MFCC y M-RASTA-PLP

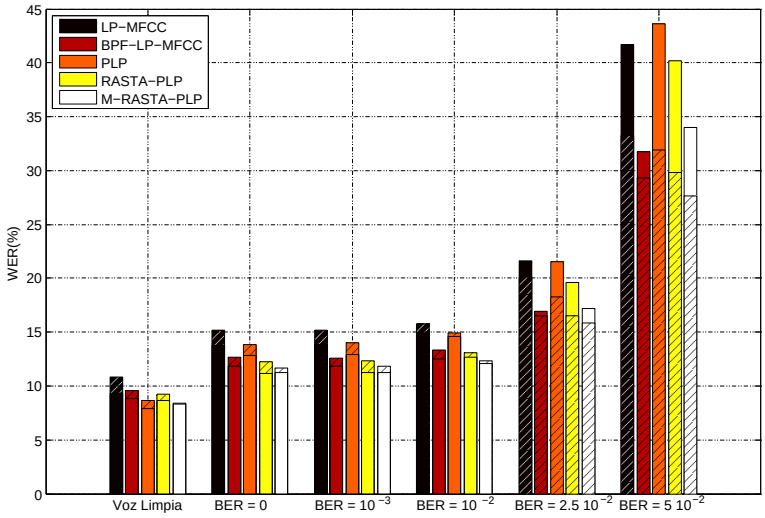


Figura 5.14: WER: Resultados aplicando BD-HMM en un entorno inalámbrico. La textura lisa hace referencia a los resultados con el reconocedor convencional y, la textura rayada, a los resultados con BD-HMM.

sistemas presentados en el Capítulo 3, BD-HMM produce peores resultados con voz limpia. No obstante, las diferencias son estadísticamente significativas únicamente para la parametrización LP-MFCC. A medida que la BER del canal empeora estas pérdidas se reducen y, finalmente, para los dos canales con mayor BER, BD-HMM

consigue resultados claramente superiores. Para los canales con baja BER, las diferencias entre aplicar BD-HMM y no hacerlo no son estadísticamente significativas, sin embargo, para canales con alta BER, generalmente encontramos significación estadística y, en concreto, para el peor canal, ésta ocurre para todas las parametrizaciones a estudio. Por otro lado, reducir el impacto de los *outliers* es especialmente importante para las parametrizaciones deducidas a partir de un análisis PLP. Así, M-RASTA-PLP conseguía peores resultados que BPF-LP-MFCC para los canales con alta BER utilizando el reconocedor convencional, sin embargo, esta situación se invierte cuando aplicamos la técnica BD-HMM. En cualquier caso, las diferencias no son estadísticamente significativas para ninguno de los canales GSM a estudio.

5.5. Conclusiones

A lo largo de este capítulo hemos estudiado el problema del reconocimiento de habla en entornos inalámbricos, prestando especial atención a la distorsión de codificación y a las distorsiones introducidas por los errores de transmisión. Las propuestas realizadas se inspiran en otros trabajos que filtran el espectro de modulación de la evolución temporal de los parámetros para mejorar la robustez de los sistemas. Hasta ahora estas técnicas de filtrado mostraban su efectividad ante distorsiones convolutivas pero, a lo largo de este capítulo, hemos visto cómo adaptarlas para combatir distorsiones propias de entornos inalámbricos.

En primer lugar hemos estudiado el efecto de eliminar las altas frecuencias que aparecen en el espectro de modulación debido a los errores de transmisión. Así, constatamos que eliminar estas altas frecuencias tanto para la parametrización MFCC como para la parametrización LP-MFCC aumenta de forma significativa las tasas de reconocimiento. Además, concluimos que la parametrización LP-MFCC presenta un comportamiento más robusto que la parametrización MFCC ante errores de transmisión. Del mismo modo, la versión paso-bajo de la parametrización LP-MFCC, es decir, la parametrización LPF-LP-MFCC, es más robusta que la parametrización que

filtra paso-bajo los coeficientes MFCC (LPF-MFCC).

Para comparar nuestras propuestas con otros trabajos previos evaluamos el método RASTA-PLP, que también se basa en el filtrado del espectro de modulación de los parámetros, en una comunicación inalámbrica. Los resultados conseguidos por RASTA-PLP en comparación con LPF-LP-MFCC nos permiten sacar dos conclusiones: 1) la sección paso-alto del filtro paso-banda del método RASTA-PLP mejora las tasas de reconocimiento en presencia de distorsiones de codificación y de errores de transmisión; y 2) la sección paso-bajo del filtro no es suficientemente selectiva para tratar adecuadamente este tipo de entornos, especialmente con canales con BERs medias o altas.

Motivados por esta conclusión, diseñamos un filtro paso-banda que combina la sección paso-alto del filtro que emplea RASTA-PLP y la sección paso-bajo que empleamos para filtrar la parametrización LP-MFCC. Aplicamos este filtro en dos entornos: 1) como alternativa al filtro paso-bajo empleado para filtrar la parametrización LP-MFCC, que denominamos BPF-LP-MFCC; y 2) como alternativa al filtro paso-banda usado en RASTA-PLP, que denominamos M-RASTA-PLP.

Los resultados experimentales indican que este nuevo filtro paso-banda proporciona mejores resultados que los filtros previos. En particular, M-RASTA-PLP mejora los resultados conseguidos por RASTA-PLP para prácticamente todas las condiciones evaluadas y, BPF-LP-MFCC es siempre mejor que LPF-LP-MFCC. En ambos casos, los incrementos son estadísticamente significativos para los canales con mayores BERs.

A continuación comparamos nuestras dos mejores soluciones: M-RASTA-PLP y BPF-LP-MFCC. Aunque ambas propuestas alcanzan resultados similares, parece claro que M-RASTA-PLP es mejor para canales con BER bajas mientras que, BPF-LP-MFCC lo es para canales con BER altas. Cuando modificamos el reconocedor aplicando la técnica BD-HMM la situación para los canales con BER altas cambia y M-RASTA-PLP consigue los mejores resultados. Esto se debe a que BD-HMM limita el efecto de los *outliers* en el reconocedor que resultan especialmente sensibles para

las parametrizaciones obtenidas a partir de un análisis PLP.

A partir de este trabajo nos surgen cuatro líneas futuras de investigación. En primer lugar nos gustaría extenderlo al codificador AMR. En segundo lugar, pretendemos evaluar el filtro propuesto cuando además de las distorsiones introducidas por el canal de comunicación nos encontramos con ruido aditivo. En tercer lugar, debemos realizar un estudio más extenso de los filtros IIR. Por último, debemos seguir investigando el filtrado del espectro de modulación cuando empleamos filtros diferentes para cada coeficiente.

Capítulo 6

Conclusiones y líneas de trabajo futuras

Las propuestas realizadas en esta tesis doctoral mejoran la robustez de los sistemas de reconocimiento automático de habla (RAH). En primer lugar, se propusieron métodos robustos ante ruidos aditivos y, a continuación, nuestras aportaciones se centraron en el tratamiento de las distorsiones introducidas por los sistemas de comunicación inalámbricos. En este capítulo resumimos las contribuciones principales que aporta esta tesis, presentamos sus conclusiones más destacables y describimos las líneas de trabajo futuras que surgen a raíz de ella.

6.1. Contribuciones

Las aportaciones principales de esta tesis pueden resumirse a través de los siguientes puntos:

- se ha puesto de manifiesto que las características muy alejadas de las medias consideradas por los modelos acústicos (*outliers*) tienen un peso excesivo en el proceso de decodificación y se propone acotar la medida de la distancia de

modo que se minimice el efecto de los *outliers*, dando lugar al método que denominamos BD-HMM.

- se sugiere combinar BD-HMM con sustracción espectral y se razona cómo cada método contrarresta las debilidades del anterior dando lugar a una sinergia que demostramos experimentalmente.

- la decodificación con incertidumbre se interpreta en el marco de las técnicas basadas en las características más fiables (*missing features*) y se obtienen expresiones analíticas para la regla de decisión de los reconocedores basados en los parámetros FF y modelos Gausianos y Uniformes de la incertidumbre.

- se demuestra que, para el caso particular de la parametrización FF, adaptar la varianza puede interpretarse como un método de decodificación con incertidumbre.

- se demuestra que, para el caso particular de la parametrización FF, el método de adaptación de varianza puede interpretarse como una ponderación de las características espectrales.

- se muestra que filtrar paso-banda el espectro de modulación resulta eficaz para reconocimiento de habla en entornos de comunicaciones inalámbricas.

6.2. Conclusiones

La primera propuesta que compone esta tesis doctoral consiste en la combinación de los métodos *bounded-distance HMM* (BD-HMM) y sustracción espectral (SS) que denominamos SSBD-HMM. [Matsui and Furui, 1992] emplean BD-HMM para mejorar las prestaciones de un sistema de reconocimiento de locutor basado en HMM y nosotros lo empleamos con el objetivo de mejorar la robustez en los sistemas RAH. BD-HMM limita el efecto que tienen los *outliers* en el proceso de reconocimiento pero, al solo actuar sobre estas observaciones, resulta adecuado combinarlo con SS,

capaz de compensar todos los parámetros. Por otro lado, SS tiene el inconveniente de introducir distorsiones en la señal de voz y, tal y como mostramos en esta tesis, uno de los efectos de estas distorsiones es que el número de *outliers* aumenta. Afortunadamente, el impacto de estos *outliers* es controlado mediante BD-HMM. De este modo, BD-HMM y SS se complementan mutuamente y, según indican nuestros resultados experimentales, su combinación tiene un efecto sinérgico.

Un método similar a BD-HMM se propuso en el entorno de reconocimiento automático de habla bajo la denominación de *acoustic backing-off* [de Veth et al., 2001a]. Este método obtuvo buenos resultados para ruidos de banda estrecha pero no resultó tan efectivo cuando los ruidos eran de banda ancha [de Veth et al., 2001b]. La combinación propuesta en esta tesis supera esta limitación y aporta mejoras significativas incluso para estos tipos de ruido.

Adicionalmente, la combinación propuesta se justificó experimentalmente cuantificando el porcentaje de muestras *outliers* que se detectan durante la etapa de reconocimiento en presencia de ruidos aditivos. Además, medimos la contribución de estos *outliers* en la regla de decisión que se emplea en el reconocimiento de habla basado en HMMs. Nuestros resultados indicaron que, aunque el porcentaje de *outliers* es bajo (alrededor de un 1-2 %), su contribución en el proceso de decodificación es muy significativa y la cuantificábamos en un entorno del 20-40 % (dependiendo de la SNR y del ruido en cuestión). A raíz de estos resultados concluimos que limitar el efecto de los *outliers* es necesario ya que, a pesar de que no contienen información relevante sobre el mensaje contenido en la señal de voz, tienen un gran peso en las decisiones que toman los reconocedores convencionales. Cuando se emplea sustracción espectral el porcentaje de *outliers* aumenta y hace este aspecto todavía más relevante.

La segunda propuesta que realizamos se basó en lo que se denomina decodificación con incertidumbre (*uncertainty decoding*). Las técnicas de regeneración de parámetros (*feature enhancement*) estiman los parámetros limpios a partir de los contaminados. Una vez realizada esta estimación, se procede con el reconocimiento

convencional. De este modo, en el proceso de decodificación no se tiene en cuenta si esas estimaciones son o no precisas. Las técnicas basadas en la decodificación con incertidumbre incorporan esta información en el reconocedor, así, modifican el algoritmo de reconocimiento de modo que se limita la contribución de las observaciones con un alto grado de incertidumbre. En esta tesis hemos diseñado sistemas apoyándonos en esta metodología cuando se utiliza sustracción espectral para estimar los parámetros limpios y los sistemas se construyen empleando la parametrización *Frequency Filtered* (FF) [Nadeu et al., 1995, Nadeu et al., 2001, Paliwal, 1999]. Cuando la incertidumbre se modeló mediante una distribución Gausiana, se vio que este método era equivalente a adaptar la varianza de los modelos. Además, al utilizar los parámetros FF podemos interpretar este método como una técnica de ponderación espectral donde el reconocimiento se basa en las componentes del espectro más fiables. Por otro lado, este método se combinó con SSBD-HMM: BD-HMM compensa el efecto de los *outliers* mientras que SS y la decodificación con incertidumbre se encargan del resto de las muestras que, estando contaminadas, no lo están tanto como estos *outliers*. A partir de nuestros resultados extraemos dos conclusiones: 1) el método que en mayor medida contribuye a las mejoras en las tasas de reconocimiento es SSBD-HMM; y, 2) los resultados conseguidos aplicando SSBD-HMM generalmente se mejoran incorporando información acerca de la incertidumbre de las observaciones en el proceso de reconocimiento.

Por último, en nuestra tercera propuesta estudiamos el efecto de otro tipo de distorsiones diferentes a las introducidas por los ruidos aditivos. En concreto, estudiamos el efecto que tiene un sistema de comunicación inalámbrico sobre los sistemas de RAH. A diferencia de lo ocurrido con los ruidos aditivos, las distorsiones introducidas por un sistema de comunicación inalámbrico no se modelan matemáticamente de manera sencilla. De este modo, nuestra propuesta no se enfocó a compensar los efectos de estas distorsiones sino a seleccionar las componentes del espectro de modulación más robustas. Así, diseñamos un filtro paso banda que seleccionase dichas componentes.

En la tesis, estudiamos el efecto de cada sección del filtro de forma separada. Así, comenzamos con la sección paso-bajo y cuantificamos las mejoras de filtrar paso-bajo la evolución temporal de los parámetros MFCC y LP-MFCC. Estos primeros resultados además de mostrar que las secuencias filtradas mejoraban los resultados conseguidos por las secuencias sin filtrar, mostraron que la parametrización LP-MFCC era la que conseguía los mejores resultados.

A continuación, cuantificamos las mejoras conseguidas al añadir una sección paso-alto de modo que filtrásemos las secuencias utilizando un filtro paso-banda. Comenzamos estudiando las prestaciones del filtro RASTA-PLP [Hermansky and Morgan, 1994] y lo comparamos con el filtro paso-bajo que habíamos diseñado. Así, se extrajeron las siguientes conclusiones: 1) la sección paso-alto de RASTA-PLP mejora las tasas de reconocimiento; y 2) la sección paso-bajo del filtro RASTA-PLP no es suficientemente abrupta para este tipo de entornos, especialmente con canales con tasas de error de bit (BER, *bit error rate*) medias o altas.

Motivados por esta conclusión, se diseñó un filtro paso-banda combinando la sección paso-alto del filtro que emplea RASTA-PLP y la sección paso-bajo que se utilizó para filtrar la parametrización LP-MFCC. Aplicamos este filtro en dos entornos: 1) como alternativa al filtro paso-bajo empleado para filtrar la parametrización LP-MFCC, que denominamos BPF-LP-MFCC; y 2) como alternativa al filtro paso-banda usado en RASTA-PLP, que denominamos M-RASTA-PLP.

Nuestros resultados experimentales indicaron que este nuevo filtro paso-banda proporciona mejores resultados que los filtros previos. En particular, M-RASTA-PLP mejora los resultados conseguidos por RASTA-PLP para prácticamente todas las condiciones evaluadas y, BPF-LP-MFCC es siempre mejor que filtrar paso-bajo la secuencia de parámetros LP-MFCC.

De entre las dos soluciones propuestas, M-RASTA-PLP resultó más adecuada para canales con BER bajas mientras que, BPF-LP-MFCC, fue la mejor solución para los canales con BER altas. Además, evaluamos las prestaciones de la técnica BD-

HMM cuando se combina con estas nuevas parametrizaciones. Nuestros resultados mostraron que BD-HMM es eficaz ante canales con BER altas pero que, sin embargo, introduce ligeras pérdidas para canales con BER menores. Las mejoras conseguidas por BD-HMM son especialmente significativas para las parametrizaciones que se extraen a través de un análisis PLP lo que hace que M-RASTA-PLP sea la que obtiene mejores resultados bajo todos los canales a estudio.

En resumen, a partir de esta tesis extraemos tres conclusiones principales:

- Los reconocedores automáticos de habla deben limitar el efecto que tienen los *outliers* en el proceso de decodificación. Tratar adecuadamente estos *outliers* mejora las prestaciones de los métodos de regeneración de parámetros tales como sustracción espectral.

- Ningún método de estimación de parámetros está carente de errores e incorporar la información acerca de la incertidumbre que todavía existe en tales estimaciones hace que los sistemas de reconocimiento sean más robustos.

- Una selección adecuada de las componentes en frecuencia del espectro de modulación mejora las prestaciones de los reconocedores ante distorsiones tan difíciles de modelar como las introducidas por un entorno inalámbrico.

6.3. Lineas de trabajo futuras

A partir del desarrollo de esta tesis nos surgen varias lineas de trabajo futuro:

- BD-HMM elimina la influencia de los *outliers* en el proceso de decodificación. De este modo, únicamente un pequeño porcentaje de observaciones se ve afectado por su actuación. Los métodos que basan el reconocimiento en las características más fiables (*missing features*) [Cooke et al., 2001, Raj et al., 2004] actúan sobre un mayor número de observaciones evitando que intervengan en el proceso de reconocimiento. Aunque estas técnicas presentan el problema de

necesitar una clasificación precisa de qué muestras son fiables y cuáles no lo son, de su aplicación se extrae una conclusión clara: no se necesitan todos los puntos del espectrograma para obtener altas tasas de reconocimiento. En este sentido nos gustaría seguir investigando qué observaciones son realmente necesarias para alcanzar estas altas tasas.

- En el capítulo que trataba sobre decodificación con incertidumbre, se asumió que sustracción espectral era capaz de eliminar el efecto que tienen los ruidos aditivos sobre la media de los parámetros. Algunos resultados preliminares indican que este asunción es menos precisa de lo que habíamos esperado y, por tanto, nos gustaría emplear estimadores más potentes de modo que se reduzca el error cometido al utilizar esta suposición.

- También estamos interesados en extender nuestro trabajo a entornos inalámbricos con nuevos codificadores de voz tales como el codificador AMR.

- Por último, nos gustaría combinar todos los métodos propuestos en esta tesis. En particular nos gustaría crear un entorno de simulación más realista donde tengamos todo tipo de distorsiones, las debidas al entorno inalámbrico y las que introducen los ruidos aditivos. En este sentido debemos adaptar la idea del filtrado del espectro de modulación para trabajar con los parámetros FF.

Apéndice A

Medias y Varianzas de la componente de ruido en los parámetros dinámicos de la parametrización FF.

En la Sección 4.2.1 del Capítulo 4 estudiamos los efectos del ruido aditivo sobre los parámetros FF. Así, caracterizamos la componente de ruido que afecta a los parámetros estáticos. Sin embargo, los sistemas diseñados a lo largo de esta tesis emplean los parámetros dinámicos para completar el vector de parametrización. En este apéndice caracterizamos la componente de ruido que afecta a estos parámetros y estimamos su media y varianza.

A.1. Parámetro dinámicos de primer orden. Parámetros deltas.

Los parámetros dinámicos se calculan a partir de los estáticos empleando la siguiente relación:

$$\Delta\widehat{FF}_{t_k} = \frac{\sum_{\theta=1}^{\Theta} \theta(\widehat{FF}_{[t+\theta]_k} - \widehat{FF}_{[t-\theta]_k})}{2\sum_{\theta=1}^{\Theta}} \tag{A.1}$$

donde $\widehat{FF}_{t_k}$ representa el k-ésimo coeficiente FF en el instante de tiempo t y Θ es un parámetro que delimita la ventana de actuación de los parámetros delta. En los experimentos realizados en esta tesis hemos usado una ventana temporal que abarca dos muestras pasadas y dos futuras, así, Θ toma el valor 2. Con este valor, reescribimos la ecuación (A.1):

$$\Delta\widehat{FF}_{t_k} = \frac{\widehat{FF}_{[t+1]_k} - \widehat{FF}_{[t-1]_k}}{10} + \frac{\widehat{FF}_{[t+2]_k} - \widehat{FF}_{[t-2]_k}}{5} \tag{A.2}$$

Para tener en cuenta el índice temporal t, reescribimos la ecuación (4.7) que relaciona la estimación de los parámetros FF, $\widehat{FF}_{t_k}$, con los parámetros limpios, FF_{t_k}:

$$\widehat{FF}_{t_k} \approx FF_{t_k} + \frac{n_{t_{(k+1)}}}{a_t} - \frac{n_{t_{(k-1)}}}{b_t} = FF_{t_k} + N_{t_k} \tag{A.3}$$

Así, usando (A.3), obtenemos la relación entre los coeficiente deltas contaminados y los que provienen de la voz limpia:

$$\Delta\widehat{FF}_{t_k} = \Delta FF_{t_k} + \frac{N_{[t+1]_t} - N_{[t-1]_k}}{10} + \frac{N_{[t+2]_k} - N_{[t-2]_k}}{5} = \Delta FF_{t_k} + \Delta N_{t_k} \tag{A.4}$$

siendo

$$\Delta N_{t_k} = \frac{N_{[t+1]_k} - N_{[t-1]_k}}{10} + \frac{N_{[t+2]_k} - N_{[t-2]_k}}{5}. \tag{A.5}$$

Las ecuaciones (4.10) y (4.11) expresan el valor de la media y varianza de N_{t_k}. Si las reescribiéndolas para tener en cuenta el índice temporal,

$$\mu_{N_{t_k}} = 0 \tag{A.6}$$

$$\sigma^2_{N_{t_k}} = \frac{\sigma^2_{n_{t_{(k+1)}}}}{a_t^2} + \frac{\sigma^2_{n_{t_{(k-1)}}}}{b_t^2}, \tag{A.7}$$

es inmediato calcular la media y la varianza de la componente de ruido de los parámetros delta:

$$\mu_{\Delta N_{t_k}} = 0 \tag{A.8}$$

$$\sigma^2_{\Delta N_{t_k}} = \frac{1}{100}\left[\sigma^2_{N_{[t+1]_k}} + \sigma^2_{N_{[t-1]_k}}\right] + \frac{1}{25}\left[\sigma^2_{N_{[t+2]_k}} + \sigma^2_{N_{[t-2]_k}}\right], \tag{A.9}$$

donde hemos supuesto que las componentes de ruido en distintos instantes de tiempo son independientes.

A.2. Parámetros dinámicos de segundo orden. Parámetros aceleración.

Los parámetros aceleración se calculan empleando la ecuación (A.1) pero sustituyendo los parámetros estáticos, $\widehat{FF}_{t_k}$, por los parámetros deltas, $\widehat{\Delta FF}_{t_k}$. La relación de los parámetros delta y estáticos es lineal y, por otro lado, la de los parámetros aceleración con los delta también es lineal. Por tanto, la relación entre los parámetros aceleración y los estáticos es lineal. Es sencillo ver que está relación viene dada por la siguiente ecuación:

$$\begin{aligned}\Delta\Delta\widehat{FF}_{t_k} = &\frac{\widehat{FF}_{[t+4]_k}}{25} + \frac{\widehat{FF}_{[t+3]_k}}{25} + \frac{\widehat{FF}_{[t+2]_k}}{100} - \frac{\widehat{FF}_{[t+1]_k}}{25} - \\ &- \frac{\widehat{FF}_{t_k}}{10} - \\ &- \frac{\widehat{FF}_{[t-1]_k}}{25} + \frac{\widehat{FF}_{[t-2]_k}}{100} + \frac{\widehat{FF}_{[t-3]_k}}{25} + \frac{\widehat{FF}_{[t-4]_k}}{25}\end{aligned} \tag{A.10}$$

Al igual que sucede con los parámetros estáticos y con los deltas, es inmediato encontrar una relación entre los parámetros aceleración contaminados y los limpios:

$$\Delta\Delta\widehat{FF}_{t_k} = \Delta\Delta FF_{t_k} + \Delta\Delta N_{t_k} \tag{A.11}$$

siendo

$$\begin{aligned}\Delta\Delta N_{t_k} = & \frac{N_{[t+4]_k}}{25} + \frac{N_{[t+3]_k}}{25} + \frac{N_{[t+2]_k}}{100} - \frac{N_{[t+1]_k}}{25} - \\ & - \frac{N_{t_k}}{10} - \\ & - \frac{N_{[t-1]_k}}{25} + \frac{N_{[t-2]_k}}{100} + \frac{N_{[t-3]_k}}{25} + \frac{N_{[t-4]_k}}{25}\end{aligned} \tag{A.12}$$

Asumiendo de nuevo que las componentes de ruido en distintos instantes de tiempo son independientes y usando las ecuaciones (A.6) y (A.7) calculamos la media y la varianza de la componente de ruido en los parámetros aceleración.

$$\mu_{\Delta\Delta N_{t_k}} = 0 \tag{A.13}$$

$$\begin{aligned}\sigma^2_{\Delta\Delta N_{t_k}} = & \frac{\sigma^2_{N_{[t+4]_k}}}{625} + \frac{\sigma^2_{N_{[t+3]_k}}}{625} + \frac{\sigma^2_{N_{[t+2]_k}}}{10^4} - \frac{\sigma^2_{N_{[t+1]_k}}}{625} - \\ & - \frac{\sigma^2_{N_{t_k}}}{100} - \\ & - \frac{\sigma^2_{N_{[t-1]_k}}}{625} + \frac{\sigma^2_{N_{[t-2]_k}}}{10^4} + \frac{\sigma^2_{N_{[t-3]_k}}}{625} + \frac{\sigma^2_{N_{[t-4]_k}}}{625}\end{aligned} \tag{A.14}$$

Apéndice B

Conclusions and future lines of research

In this Ph.D. Thesis several strategies have been suggested and studied to improve the robustness of current automatic speech recognition (ASR) systems. We started proposing methods for improving the robustness to additive noises and we followed proposing methods that mitigate the effects of modern wireless communications environments. Here, we present a summary of our main conclusions. Next, we present the lines of research that remain open and, in our opinion, deserve future attention.

B.1. Conclusions

Our first proposal was a combination of bounded-distance HMM (BD-HMM) and spectral subtraction (SS), that we called SSBD-HMM. BD-HMM was first applied in HMM-based speaker recognition [Matsui and Furui, 1992] and we used it for robust ASR. BD-HMM limits the impact of outlier features on the recognition process. However, since BD-HMM just deals with outliers it turns out appropriate to combine it with SS, that enhances all the features. As we experimentally proved, the side

effect of SS was that the number of outliers increased. Fortunately, these outliers were properly countered by BD-HMM. As a result, SS and BD-HMM complemented each other well and, from our results, we conclude that this combination exhibits a synergy effect able to notably improve the recognition performance.

Acoustic backing-off [de Veth et al., 2001a] is a method similar to BD-HMM proposed for speech recognition. This method obtained good results for narrow-band noises but the performance was poor for wide-band noises [de Veth et al., 2001b]. The proposed combination, SSBD-HMM, overcame these limitations and achieved significant improvements for wide-band noises.

Additionally, the proposed combination was motivated by accounting the actual percentage of outliers in presence of additive noises for a well-known ASR task. Furthermore, we measured their contribution in the recognition process. Although the percentage was low (around $1-2$ %), their influence on the recogniser was severe (their contribution to the accumulated log-likelihood was around $20-40$ %). Consequently, we conclude that limiting the outlier contribution is necessary because, although they do not contain relevant information about the embedded message, they have a strong weight on the decisions taken by conventional recognisers. When SS was applied, the percentage of outliers increased making this issue even more important.

Our second approach was based on uncertainty decoding. Parameter enhancement methods aim at estimating the clean parameters from the noisy ones. Once the estimated parameters are available, the recognition process typically continues without taking into consideration the quality of those estimates. Uncertainty decoding methods take into consideration some information on the quality of the features. They modify the decoding algorithm in a way that features that exhibit a high degree of uncertainty do not play a relevant role in the recognition. In this Thesis we built a system based on these ideas that used the FF parameterization [Nadeu et al., 1995, Nadeu et al., 2001, Paliwal, 1999] and SS for feature enhancement. When the uncertainty was modelled by means of a Gaussian distribution, we showed that this method became equivalent to adapt the model variances. Furt-

hermore, the use of FF parameters allowed us to interpret this method as a spectral weighting that assigns more importance to the most reliable spectral components. Additionally, we combined this method with SSBD-HMM: BD-HMM dealt with outliers while SS and uncertainty decoding dealt with the remaining features. From our results, two main conclusions can be drawn: 1) SSBD-HMM is the method that contribute the most in terms of reducing WER; and 2) the results achieved by SSBD-HMM are generally improved by incorporating information about the feature uncertainty in the decoding process.

Finally, our third proposal was focused on distortions different from additive noises. Specifically, we dealt with distortions owing to wireless communication systems and we studied how the ASR performance was affected by these distortions. While modelling the effects of additive noises on the recongnition system was analytically tractable, the distortions owing to the wireless environment were more difficult to model and, therefore, our approach was not oriented towards compensating their effects. Instead, we tried to weigth the features aiming at emphasizing the least affected features and de-emphasizing the most affect ones. Specifically, we studied their effects on the feature spectra and suggested a band-pass filtering to properly weigth the frequency components of the feature spectra.

We looked into the two sections of the band-pass filter separately. We started with the low-pass section and we applied a low-pass filter to the MFCC and LP-MFCC parameterizations. First, we realised that we had different behaviours even for the unfiltered features: LP-MFCC parameterization was superior to the MFCC parameterization. Next, we observed that the low-pass filtered features were more robust than the unfiltered ones. From the two filtered sequences, the filtered LP-MFCC sequence provided the best results.

Next, we add a high-pass section to quantify its potential benefits. We began with the well known RASTA-PLP band-pass filter [Hermansky and Morgan, 1994]. We compared its performance with the low-pass filter and two conclusions were extracted: 1) the high-pass section of the RASTA-PLP band-pass filter produced improvements

in the recognition performance; and 2) the low-pass section of the RASTA-PLP filter was not sharp enough to deal with this type of distortions, especially for medium and high BER channels.

Motivated by these last conclusions, we designed a band-pass filter combining the high-pass section of the RASTA-PLP filter with the low-pass section that we had initially proposed. We applied this filter in two configurations: 1) as an alternative to the low-pass filtering proposed to filter the LP-MFCC, called BPF-LP-MFCC (Band-Pass Filtering LP-MFCC); and 2) as an alternative to the band-pass filter of RASTA-PLP, leading what we called M-RASTA-PLP (Modified RASTA-PLP). In both configurations, the new filter outperformed the former ones, especially for medium and high BER channels. In particular, M-RASTA-PLP was superior to RASTA-PLP for almost every channel condition and BPF-LP-MFCC was always better than low-pass filtering the LP-MFCC parameters.

Next, we compared M-RASTA-PLP and BPF-LP-MFCC to conclude that, although both parameter sets yielded similar results, M-RASTA-PLP was the best option for low BER channels while BPF-LP-MFCC was preferable for high BER channels. Furthermore, we evaluated the performance of BD-HMM in combination with these new parameterizations. From our results, we notice that BD-HMM is effective for high BER channels while introduces a slight loss of performance for channels with lower BERs. The improvements were especially significant for parameters extracted from a PLP analysis which makes M-RASTA-PLP the best choice for most of the conditions.

Two sum up, we would like to emphasize three main conclusions:

1. Speech recognisers should avoid the outliers to take part in the decoding process, since they do not carry discriminative information. Dealing with outliers improves the performance of speech enhancement methods such as spectral subtraction.

2. None estimator is error free and incorporating information about the uncer-

tainty that remains in the estimates makes the systems more robust.

3. A proper selection of the frequencies in the modulation spectrum improves the recognition accuracy when dealing with distortions as difficult to model as the ones produced in a wireless environments.

B.2. Future lines of research

This Thesis has left open questions that establish future lines of research. In this section we present these future lines:

- BD-HMM removes the effects of the outliers in the recognition process. Thus, this method just deals with a low percentage of features. However, missing features techniques [Cooke et al., 2001, Raj et al., 2004] avoid a larger number of features to participate in the decoding process with success. Although these methods have the problem of needing a high accurate detector that labels the features as reliable/unreliable, they conduce to a clear conclusion: not all the time-frequency components of the speech signal are needed to get competitive recognisers. In this sense, we would like to investigate which features are really needed. In particular, we think that not all the features should have the same importance in the decoding process and a proper selection of the relevant ones will probably make the systems more robust.

- When we applied methods based on uncertainty decoding, we assume that spectral subtraction was able to remove the effects of additive noises on the parameter means. Some preliminary results indicate that this assumption is not as precise as we had expected. We would like to use more powerful estimators so this assumption would be more exact.

- We would also like to extend our work about ASR and wireless environments with newer speech coders like the AMR family.

- Finally, we would like to combine all the methods that have been proposed in this Thesis. In particular we would like to create a more realistic environment including additive noises and distortions owing to wireless communication sytems. In this sense, we should adapt the idea of filtering the modulation spectrum to the FF parameters.

Bibliografía

[Acero and Stern, 1990] Acero, A. and Stern, R. (1990). Environmental robustness in automatic speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 849–852.

[Arrowood and Clements, 2002] Arrowood, J. A. and Clements, M. A. (2002). Using observation uncertainty in HMM decoding. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1561–1564.

[Barker et al., 2005] Barker, J., Cooke, M., and Ellis, D. (2005). Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25.

[Barker et al., 2001] Barker, J., Cooke, M., and Green, P. (2001). Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pages 213–217.

[Benitez et al., 2004] Benitez, C., Segura, J. C., de la Torre, A., Ramirez, J., and Rubio, A. J. (2004). Including uncertainty of speech observations in robust speech recognition. In *Proc. Int. Conf. on Spoken Language Processing (INTERSPEECH - ICSLP)*, pages 137–140.

[Bernard and Alwan, 2002] Bernard, A. and Alwan, A. (2002). Low-bitrate distributed speech recognition for packet-based and wireless communication. *IEEE Trans. Speech Audio Processing*, 10(8):570–580.

[Berouti et al., 1979] Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 4, pages 208–211.

[Boll, 1979] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoustics, Speech, Signal Processing*, 27(2):113–120.

[Carlson and Clements, 1991] Carlson, B. A. and Clements, M. A. (1991). Application of a weighted projection measure for robust hidden Markov model based speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 921–924.

[Chen et al., 2004] Chen, B., Zhu, Q., and Morgan, N. (2004). Learning long-term temporal features in LVCSR using neural networks. In *Proc. Int. Conf. on Spoken Language Processing (INTERSPEECH - ICSLP)*, pages 925–928.

[Chien et al., 1995] Chien, J.-T., Lee, L.-M., and Wang, H.-C. (1995). Noisy speech recognition by using variance adapted hidden Markov models. *Electronics Lett.*, 31(18):1555–1556.

[CMU, 1998] CMU (1998). The CMU (v 0.6) pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

[Cooke et al., 2001] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(3):267–285.

[Cui and Alwan, 2004] Cui, X. and Alwan, A. (2004). Combining feature compensation and weighted Viterbi decoding for noise robust speech recognition with limited adaptation data. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 969–972.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, Signal Processing*, 28(4):357–366.

[de la Torre et al., 2005] de la Torre, A., Peinado, A., Segura, J., Perez-Cordoba, J., Benitez, M., and Rubio, A. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Speech Audio Processing*, 13(3):355–366.

[de Veth et al., 1998] de Veth, J., Cranen, B., and Boves, L. (1998). Acoustic backing-off in the local distance computation for robust automatic speech recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1427–1430.

[de Veth et al., 2001a] de Veth, J., Cranen, B., and Boves, L. (2001a). Acoustic backing-off as an implementation of missing feature theory. *Speech Communication*, 34(3):247–265.

[de Veth et al., 2001b] de Veth, J., de Wet, F., Cranen, B., and Boves, L. (2001b). Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR. *Speech Communication*, 34(1):57–74.

[Deng et al., 2002] Deng, L., Droppo, J., and Acero, A. (2002). Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2449–2452.

[Digalakis et al., 1999] Digalakis, V., Neumeyer, L., and Perakakis, M. (1999). Quantization of cepstral parameters for speech recognition over the World Wide Web. *IEEE Journal on Selected Areas in Communications*, 17(1):82–90.

[Droppo et al., 2002] Droppo, J., Acero, A., and Deng, L. (2002). ncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 57–60.

[ETSI ES 201 108, 2003] ETSI ES 201 108 (2003). Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms. Ver. 1.1.3.

[ETSI ES 202 050, 2004] ETSI ES 202 050 (2004). Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms. Ver. 1.1.1.

[ETSI ETS 300 578, 1999] ETSI ETS 300 578 (1999). Digital cellular telecommunications system (phase 2); Radio subsystem link control (GSM 05.08 version 4.22.1).

[ETSI Recommendation GSM 6.20, 1999] ETSI Recommendation GSM 6.20 (1999). Digital cellular telecommunications systems; Half Rate speech; Part 2: Half Rate speech transcoding.

[Euler and Zinke, 1994] Euler, S. and Zinke, J. (1994). The influence of speech coding algorithms on automatic speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 621–624.

[Furui, 1981] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoustics, Speech, Signal Processing*, 29(2):254–272.

[Furui, 1986] Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoustics, Speech, Signal Processing*, 34(1):52–59.

[Gales, 1998] Gales, M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98.

[Gales and Young, 1996] Gales, M. and Young, S. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Trans. Speech Audio Processing*, 4(5):352–359.

[Gallardo-Antolin et al., 2005] Gallardo-Antolin, A., Pelaez-Moreno, C., and Diaz-de Maria, F. (2005). Recognizing GSM digital speech. *IEEE Trans. Speech Audio Processing*, 13(6):1186–1205.

[Gauvain and Lee, 1994] Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Processing*, 2(2):291–298.

[Gold and Morgan, 2000] Gold, B. and Morgan, N. (2000). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, chapter 22: Feature extraction for ASR, pages 295–308. John Wiley & Sons.

[Gong, 1995] Gong, Y. (1995). Speech recognition in noisy environments: a survey. *Speech Communication*, 16(3):261–291.

[Greenberg, 1996] Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pages 1–8.

[Hanson and Applebaum, 1993] Hanson, B. and Applebaum, T. (1993). Subband or cepstral domain filtering for recognition of lombard and channel-distorted speech. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 79–82.

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Trans. Speech Audio Processing*, 2(4):578–589.

[Hirsch, 2002a] Hirsch, G. (2002a). Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, Version 2.0, AU/417/02. Technical report, ETSI STQ-Aurora DSR Working Group.

[Hirsch, 2002b] Hirsch, H.-G. (2002b). The influence of speech coding on recognition performance in telecommunication networks. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1877–1880.

[Juang et al., 1987] Juang, B.-H., Rabiner, L., and Wilpon, J. (1987). On the use of bandpass liftering in speech recognition. *IEEE Trans. Speech Audio Processing*, 35(7):947–954.

[Junqua, 2000] Junqua, J. C. (2000). *Robust speech recognition in embedded systems and PC applications*. Kluwer Academic Publishers, 2000.

[Junqua and Haton, 1995] Junqua, J.-C. and Haton, J.-P. (1995). *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, chapter 8: On the use of a robust speech representation, pages 233–272. Kluwer Academic Publishers.

[Jurafsky and Martin, 2000] Jurafsky, D. and Martin, J. H. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall.

[Kanedera et al., 1998] Kanedera, N., Hermansky, H., and Arai, T. (1998). On properties of modulation spectrum for robust automatic speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 613–616.

[Kim et al., 2002] Kim, H. K., Cox, R., and Rose, R. (2002). Performance improvement of a bitstream-based front-end for wireless speech recognition in adverse environments. *IEEE Trans. Speech Audio Processing*, 10(8):591–604.

[Kiss et al., 2003] Kiss, I., Lakaniemi, A., Yang, C., and Viikki, O. (2003). Review of AMR speech codec-and distributed speech recognition-based speech-enabled

services. In *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, pages 613–618.

[Krisjansson and Frey, 2002] Krisjansson, T. T. and Frey, B. J. (2002). Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 61–64.

[Lilly and Paliwal, 1996] Lilly, B. and Paliwal, K. (1996). Effect of speech coders on speech recognition performance. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, volume 4, pages 2344–2347.

[Lippmann, 1997] Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

[Macho, 2000] Macho, D. (2000). Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: Description and baseline results. Technical report, UPC, Universitat Politecnica de Catalunya.

[Martin, 2001] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Processing*, 9(5):504–512.

[Matsui and Furui, 1991] Matsui, T. and Furui, S. (1991). A text-independent speaker recognition method robust against utterance variations. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 377–380.

[Matsui and Furui, 1992] Matsui, T. and Furui, S. (1992). Comparison of test-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 157–160.

[Molau et al., 2003a] Molau, S., Hilger, F., and Ney, H. (2003a). Feature space normalization in adverse acoustic conditions. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 656–659.

[Molau et al., 2003b] Molau, S., Keysers, D., and Ney, H. (2003b). Matching training and test data distributions for robust speech recognition. *Speech Communication*, 41(4):579–601.

[Moreno et al., 1996] Moreno, P., Raj, B., and Stern, R. (1996). A vector Taylor series approach for environment-independent speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 733–736.

[Morris et al., 2001] Morris, A., Barker, J., , and Bourlard, H. (2001). From missing data to maybe useful data: soft data modelling for noise robust ASR. In *WISP workshop on innovative methods in speech recognition.*

[Nadeu et al., 1995] Nadeu, C., Hernando, J., and Gorricho, M. (1995). On the decorrelation of filter-bank energies in speech recognition. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pages 1381–1384.

[Nadeu et al., 2001] Nadeu, C., Macho, D., and Hernando, J. (2001). Time and frequency filtering of filter-bank energies for robust HMM speech recognition. *Speech Communication*, 34(1-2):93–114.

[Nadeu et al., 1997] Nadeu, C., Pachès-Leal, P., and Juang, B.-H. (1997). Filtering the time sequences of spectral parameters for speech recognition. *Speech Communication*, 22(4):315–332.

[NIST, 1992] NIST (1992). NIST, The Resource Management Corpus(RM1). *Distributed by NIST.*

[Nolazco Flores and Young, 1994] Nolazco Flores, J. and Young, S. (1994). Continuous speech recognition in noise using spectral subtraction and HMM adapta-

tion. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 1, pages 409–412.

[O'Shaughnessy, 1999] O'Shaughnessy, D. (1999). *Speech Communications: Human and Machine*, chapter 6: Speech Analysis. IEEE Press; 2nd edition.

[Paliwal, 1982] Paliwal, K. K. (1982). On the performance of the quefrency-weighted cepstral coefficients in vowel recognition. *Speech Communication*, 1(2):151–154.

[Paliwal, 1999] Paliwal, K. K. (1999). Decorrelated and liftered filter-bank energies for robust speech recognition. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pages 85–88.

[Papoulis and Pillai, 2002] Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variable and Stochastic Processes*. McGraw-Hill, 4th edition.

[Paul and Baker, 1992] Paul, D. B. and Baker, J. M. (1992). The design for the wall street journal-based CSR corpus. In *Human Language Technology Conference*, pages 357–362.

[Pelaez-Moreno et al., 2001] Pelaez-Moreno, C., Gallardo-Antolin, A., and Diaz-de Maria, F. (2001). Recognizing voice over IP: a robust front-end for speech recognition on the World Wide Web. *IEEE Transactions on Multimedia*, 3(2):209–218.

[Peláez-Moreno et al., 2002] Peláez-Moreno, C., Gallardo-Antolín, A., Vicente-Peña, J., and de María, F. D. (2002). Filtering the spectral parameters to mitigate the influence of transmission errors on ASR systems. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2217–2220.

[Pujol et al., 2004] Pujol, P., Nadeu, C., Macho, D., and Padrell, J. (2004). Speech recognition experiments with the SPEECON database using several robust front-ends. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286.

[Raj, 2000] Raj, B. (2000). *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*. PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

[Raj et al., 2004] Raj, B., Seltzer, M., and Stern, R. (2004). Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43(4):275–296.

[Raj and Stern, 2005] Raj, B. and Stern, R. (2005). Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, 22(5):101–116.

[Seltzer et al., 2004] Seltzer, M., Raj, B., and Stern, R. (2004). A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43(4):379–393.

[Shozakai et al., 1997] Shozakai, M., Nakamura, S., and Shikano, K. (1997). A speech enhancement approach E-CMN/CSS for speech recognition in car environments. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 450–457.

[Smolders and Van Compernolle, 1993] Smolders, J. and Van Compernolle, D. (1993). In search for the relevant parameters for speaker independent speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 2, pages 684–687.

[Soong and Sondhi, 1988] Soong, F. and Sondhi, M. (1988). A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise. *IEEE Trans. Acoustics, Speech, Signal Processing*, 36(1):41–48.

[Stahl et al., 2000] Stahl, V., Fischer, A., and Bippus, R. (2000). Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, volume 3, pages 1875–1878.

[Stern et al., 1996] Stern, R. M., Acero, A., Liu, F.-H., and Ohshima, Y. (1996). *Speech Recognition*, chapter Signal Processing for Robust Speech Recognition, pages 351–378. Kluwer Academic Publishers.

[Stern et al., 1997] Stern, R. M., Raj, B., and Moreno, P. J. (1997). Compensation for environmental degradation in automatic speech recognition. In *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 33–42.

[Stouten et al., 2006] Stouten, V., hamme, H. V., and Wambacq, P. (2006). Model-based feature enhancement with uncertainty decoding for noise robust ASR. *Speech Communication*, 48(11):1502–1514.

[Tohkura, 1987] Tohkura, Y. (1987). A weighted cepstral distance measure for speech recognition. *IEEE Trans. Acoustics, Speech, Signal Processing*, 35(10):1414–1422.

[Varga et al., 1992] Varga, A. P., Steenneken, J. M., Tomlinson, M., and Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on Automatic Speech Recognition. In *Tech. Rep. DRA Speech Res. Unit. Malvern, Worcestershire, U. K.*

[Vicente-Peña et al., 2006a] Vicente-Peña, J., Díaz-de-María, F., and Kleijn, W. B. (2006a). Individual on-line variance adaptation of frequency filtered parameters for robust ASR. In *Proc. Int. Conf. on Spoken Language Processing (INTERSPEECH - ICSLP)*, pages 1491–1494.

[Vicente-Peña et al., 2007] Vicente-Peña, J., Díaz-de-María, F., and Kleijn, W. B. (2007). The synergy between bounded-distance HMM and spectral subtraction for robust speech recognition. *Submitted to Speech Communication.*

[Vicente-Peña et al., 2006b] Vicente-Peña, J., Gallardo-Antolín, A., Peláez-Moreno, C., and de María, F. D. (2006b). Band-pass filtering of the time sequences of

spectral parameters for robust wireless speech recognition. *Speech Communication*, 48(10):1379–1398.

[Viikki and Laurila, 1998] Viikki, O. and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1):133–147.

[Weiss and Hasset, 1993] Weiss, N. A. and Hasset, M. J. (1993). *Introductory statistics*, pages 407–408. Addison-Wesley, third edition.

[Yoma et al., 1998] Yoma, N., McInnes, F., and Jack, M. (1998). Improving performance of spectral subtraction in speech recognitionusing a model for additive noise. *IEEE Trans. Speech Audio Processing*, 6(6):579–582.

[Yoma et al., 1995] Yoma, N. B., McInnes, F., and Jack, M. (1995). Improved algorithms for speech recognition in noise using lateral inhibition and SNR weighted. In *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, pages 461–464.

[Young et al., 2002] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book (for HTK Version 3.2.1)*. Cambridge Univ. Press, Cambridge, U.K.